

## ESTIMATION OF THE TRANSITION DENSITY OF A MARKOV CHAIN

MATHIEU SART

ABSTRACT. We present two data-driven procedures to estimate the transition density of an homogeneous Markov chain. The first yields to a piecewise constant estimator on a suitable random partition. By using an Hellinger-type loss, we establish non-asymptotic risk bounds for our estimator when the square root of the transition density belongs to possibly inhomogeneous Besov spaces with possibly small regularity index. Some simulations are also provided. The second procedure is of theoretical interest and leads to a general model selection theorem from which we derive rates of convergence over a very wide range of possibly inhomogeneous and anisotropic Besov spaces. We also investigate the rates that can be achieved under structural assumptions on the transition density.

## 1. INTRODUCTION.

Consider a time-homogeneous Markov chain  $(X_i)_{i \in \mathbb{N}}$  defined on an abstract probability space  $(\Omega, \mathcal{E}, \mathbb{P})$  with values in the measured space  $(\mathbb{X}, \mathcal{F}, \mu)$ . We assume that for each  $x \in \mathbb{X}$ , the conditional law  $\mathcal{L}(X_{i+1} \mid X_i = x)$  admits a density  $s(x, \cdot)$  with respect to  $\mu$ . Our aim is to estimate the transition density  $(x, y) \mapsto s(x, y)$  on a subset  $A = A_1 \times A_2$  of  $\mathbb{X}^2$  from the observations  $X_0, \dots, X_n$ .

Many papers are devoted to this statistical setting. A popular method to build an estimator of  $s$  is to divide an estimator of the joint density of  $(X_i, X_{i+1})$  by an estimator of the density of  $X_i$ . The resulting estimator is called a quotient estimator. Roussas (1969), Athreya and Atun-car (1998) considered Kernel estimators for the densities of  $X_i$  and  $(X_i, X_{i+1})$ . They proved consistence and asymptotic normality of the quotient estimator. Other properties of this estimator were established: Roussas (1991), Dorea (2002) showed strong consistency, Basu and Sahoo (1998) proved a Berry-Essen type theorem and Doukhan and Ghindès (1983) bounded from above the integrated quadratic risk under Sobolev constraints. Cléménçon (2000) investigated the minimax rates when  $A = [0, 1]^2$ ,  $\mathbb{X}^2 = \mathbb{R}^2$ . Given two smoothness classes  $\mathcal{F}_1$  and  $\mathcal{F}_2$  of real valued functions on  $[0, 1]^2$  and  $[0, 1]$  respectively (balls of Besov spaces), he established the lower bounds over the class

$$\mathcal{F} = \left\{ \varphi, \forall x, y \in [0, 1], \varphi(x, y) = \frac{\varphi_1(x, y)}{\varphi_2(x)}, (\varphi_1, \varphi_2) \in \mathcal{F}_1 \times \mathcal{F}_2 \right\}.$$

He developed a method based on wavelet thresholding to estimate the densities of  $X_i$  and  $(X_i, X_{i+1})$  and showed that the quotient estimator of  $s$  is quasi-optimal in the sense that the

---

*Date:* October, 2012.

*2010 Mathematics Subject Classification.* 62M05, 62G05.

*Key words and phrases.* Adaptive estimation, Markov chain, Model selection, Robust tests, Transition density.

minimax rates are achieved up to possible logarithmic factors. Lacour (2008) used model selection via penalization to construct estimates of the densities. The resulting quotient estimator reaches the minimax rates over  $\mathcal{F}$  when  $\mathcal{F}_1$  and  $\mathcal{F}_2$  are balls of homogeneous (but possibly anisotropic) Besov spaces on  $[0, 1]^2$  and  $[0, 1]$  respectively.

The previous rates of convergence depend on the smoothness properties of the densities of  $X_i$  and  $(X_i, X_{i+1})$ . In the favourable case where  $X_0, \dots, X_n$  are drawn from a stationary Markov chain (with stationary density  $f$ ), the rates depend on the smoothness properties of  $f$  or more precisely on the restriction of  $f$  to  $A_1$ . This function may however be less regular than the target function  $s$ . We refer for instance to Section 5.4.1 of Cl  mencon (2000) for an example of a Doeblin recurrent Markov chain where the stationary density  $f$  is discontinuous on  $[0, 1]$  although  $s$  is constant on  $[0, 1]^2$ . Therefore, these estimators may converge slowly even if  $s$  is smooth, which is problematic.

This issue was overcome in several papers. Cl  mencon (2000) proposed a second procedure, based on wavelets and an analogy with the regression setting. He computed the lower bounds of minimax rates when the restriction of  $s$  on  $[0, 1]^2$  belongs to balls of some (possibly inhomogeneous) Besov spaces and proved that its estimator achieves these rates up to a possible logarithmic factor. Lacour (2007) established lower bound over balls of some (homogeneous but possibly anisotropic) Besov spaces. By minimizing a penalized contrast inspired from the least-squares, she obtained a model selection theorem from which she deduced that her estimator reaches the minimax rates when  $A = [0, 1]^2$ ,  $\mathbb{X}^2 = \mathbb{R}^2$ . With a similar procedure, Akakpo and Lacour (2011) obtained the usual rates of convergence over balls of possibly anisotropic and inhomogeneous Besov spaces (when  $\mathbb{X}^2 = A = [0, 1]^{2d}$ ). Very recently, Birg   (2012) proposed a procedure based on robust testing to establish a general oracle inequality. The expected rates of convergence can be deduced from this inequality when  $\sqrt{s}$  belongs to balls of possibly anisotropic and inhomogeneous Besov spaces.

These authors have used different losses in order to evaluate the performance of their estimators. In each of these papers, the risk of an estimator  $\hat{s}$  is of the form  $\mathbb{E} [\delta^2(s \mathbb{1}_A, \hat{s})]$  where  $\mathbb{1}_A$  denotes the indicator function of the subset  $A$  and  $\delta$  a suitable distance. Lacour (2007), Akakpo and Lacour (2011) considered the space  $\mathbb{L}^2(\mathbb{X}^2, M)$  of square integrable functions on  $\mathbb{X}^2$  equipped with the random product measure  $M = \lambda_n \otimes \mu$  where  $\lambda_n = n^{-1} \sum_{i=0}^{n-1} \delta_{X_i}$  and used the distance defined for  $f, f' \in \mathbb{L}^2(\mathbb{X}^2, M)$  by

$$\delta^2(f, f') = \frac{1}{n} \sum_{i=0}^{n-1} \int_{\mathbb{X}} (f(X_i, y) - f'(X_i, y))^2 d\mu(y).$$

Birg   (2012) considered the cone  $\mathbb{L}_+^1(\mathbb{X}^2, \mu \otimes \mu)$  of non-negative integrable functions and used the deterministic Hellinger-type distance defined for  $f, f' \in \mathbb{L}_+^1(\mathbb{X}^2, \mu \otimes \mu)$  by

$$\delta^2(f, f') = \frac{1}{2} \int_{\mathbb{X}^2} \left( \sqrt{f(x, y)} - \sqrt{f'(x, y)} \right)^2 d\mu(x) d\mu(y).$$

These approaches, which often rely on the loss that is used, require the knowledge (or at least a suitable estimation) of various quantities depending on the unknown  $s$ , such as the supremum norm of  $s$ , or on a positive lower bound, either on the stationary density, or on  $k^{-1} \sum_{j=1}^k s^{(l+j)}$  for some  $k \geq 1$ ,  $l \geq 0$  where  $s^{(l+j)}(x, \cdot)$  is the density of the conditional law  $\mathcal{L}(X_{l+j} \mid X_0 = x)$ . Unfortunately, these quantities not only influence the way the estimators are built but also their

performances since they are involved in the risk bounds. In the present paper, we shall rather consider the distance  $H$  (corresponding to an analogue of the random  $\mathbb{L}^2$  loss above) defined on the cone  $\mathbb{L}_+^1(\mathbb{X}^2, M)$  of integrable and non-negative functions by

$$H^2(f, f') = \frac{1}{2n} \sum_{i=0}^{n-1} \int_{\mathbb{X}} \left( \sqrt{f(X_i, y)} - \sqrt{f'(X_i, y)} \right)^2 d\mu(y) \quad \text{for all } f, f' \in \mathbb{L}_+^1(\mathbb{X}^2, M).$$

For such a loss, we shall show that our estimators satisfy an oracle-type inequality under very weak assumptions on the Markov chain. A connection with the usual deterministic Hellinger-type loss will be done under a posteriori assumptions on the chain, and hence, independently of the construction of the estimator.

Our estimation strategy can be viewed as a mix between an approach based on the minimization of a contrast and an approach based on robust tests. Estimation procedures based on tests started in the seventies with Lucien Lecam and Lucien Birgé (LeCam (1973, 1975); Birgé (1983); Birgé (1984a,b)). More recently, Birgé (2006) presented a powerful device to establish general oracle inequalities from robust tests. It was used in our statistical setting in Birgé (2012) and in many others in Birgé (2007, 2008) and Sart (2012). We make two contributions to this area. Firstly, we provide a new test for our statistical setting. This test is based on a variational formula inspired from Baraud (2010) and differs from the one of Birgé (2012). Secondly, we shall study procedures that are quite far from the original one of Birgé (2006). Let us explain why.

The procedure of Birgé (2006) depends on a suitable net, the construction of which is usually abstract, making thus the estimator impossible to build in practice. In the favourable cases where the net can be made explicit, the procedure is anyway too complex to be implemented (see for instance Section 3.4.2 of Birgé (2007)). This procedure was afterwards adapted to estimators selection in Baraud and Birgé (2009) (for histogram type estimators) and in Baraud (2010) (for more general estimators). The complexity of their algorithms is of order the square of the cardinality of the family and are thus implementable when this family is not too large. In particular, given a family of histogram type estimators  $\{\hat{s}_m, m \in \mathcal{M}\}$ , these two procedures are interesting in practice when  $\mathcal{M}$  is a collection of regular partitions (namely when all its elements have same Lebesgue measure) but become unfortunately numerically intractable for richer collections. In this work, we tackle this issue by proposing a new way of selecting among a family of piecewise constant estimators when the collection  $\mathcal{M}$  ensues from the adaptive approximation algorithm of DeVore and Yu (1990).

We present this procedure in the first part of the paper. It yields to a piecewise constant estimator on a data-driven partition that satisfies an oracle-type inequality from which we shall deduce uniform rates of convergence over balls of (possibly) inhomogeneous Besov spaces with small regularity indices. These rates coincide, up to a possible logarithmic factor to the usual ones over such classes. Finally, we carry out numerical simulations to compare our estimator with the one of Akakpo and Lacour (2011).

In the second part of this paper, we are interested in obtaining stronger theoretical results for our statistical problem. We put aside the practical considerations to focus on the construction of an estimator that satisfies a general model selection theorem. Such an estimator should be considered as a benchmark for what theoretically feasible. We deduce rates of convergence over a large range of anisotropic and inhomogeneous Besov spaces on  $[0, 1]^{2d}$ . We shall also consider

other kinds of assumptions on the transition density. We shall assume that  $s$  belongs to classes of functions satisfying structural assumptions and for which faster rates of convergence can be achieved. This approach was developed by Juditsky et al. (2009) (in the Gaussian white noise model) and by Baraud and Birgé (2011) (in more statistical settings) to avoid the curse of dimensionality. More precisely, Baraud and Birgé (2011) showed that these rates can be deduced from a general model selection theorem, which strengthens its theoretical interest. This strategy was used in Sart (2012) to establish risk bounds over many classes of functions for Poisson processes with covariates. We shall use these assumptions to obtain faster rates of convergence for autoregressive Markov chains (whose conditional variance may not be constant).

This paper is organized as follows. The first procedure, which selects among piecewise constant estimators is presented and theoretically studied in Section 2. In Section 3, we carry out a simulation study and compare our estimator with the one of Akakpo and Lacour (2011). The practical implementation of this procedure is quite technical and will therefore be delayed in the appendix, in Section 5. In Section 4, we establish theoretical results by using our second procedure. The proofs are postponed to Section 6.

Let us introduce some notations that will be used all along the paper. The number  $x \vee y$  (respectively  $x \wedge y$ ) stands for  $\max(x, y)$  (respectively  $\min(x, y)$ ) and  $x_+$  stands for  $x \vee 0$ . We set  $\mathbb{N}^* = \mathbb{N} \setminus \{0\}$ . For  $(E, d)$  a metric space,  $x \in E$  and  $A \subset E$ , the distance between  $x$  and  $A$  is denoted by  $d(x, A) = \inf_{a \in A} d(x, a)$ . The indicator function of a subset  $A$  is denoted by  $\mathbb{1}_A$  and the restriction of a function  $f$  to  $A$  by  $f|_A$ . For all real valued function  $f$  on  $E$ ,  $\|f\|_\infty$  stands for  $\sup_{x \in E} |f(x)|$ . The cardinality of a finite set  $A$  is denoted by  $|A|$ . The notations  $C, C', C'' \dots$  are for the constants. The constants  $C, C', C'' \dots$  may change from line to line.

## 2. SELECTING AMONG PIECEWISE CONSTANT ESTIMATORS.

Throughout this section, we assume that  $\mathbb{X} = \mathbb{R}^d$ ,  $A = [0, 1]^{2d}$ ,  $\mu([0, 1]^d) = 1$  and  $n > 3$ .

**2.1. Preliminary estimators.** Given a (finite) partition  $m$  of  $[0, 1]^{2d}$ , a simple way to estimate  $s$  on  $[0, 1]^{2d}$  is to consider the piecewise constant estimator on the elements of  $m$  defined by

$$(1) \quad \hat{s}_m = \sum_{K \in m} \frac{\sum_{i=0}^{n-1} \mathbb{1}_K(X_i, X_{i+1})}{\sum_{i=0}^{n-1} \int_{[0,1]^d} \mathbb{1}_K(X_i, x) d\mu(x)} \mathbb{1}_K.$$

In the above definition, the denominator  $\sum_{i=0}^{n-1} \int_{\mathbb{X}} \mathbb{1}_K(X_i, x) d\mu(x)$  may be equal to 0 for some sets  $K$ , in which case the numerator  $\sum_{i=0}^{n-1} \mathbb{1}_K(X_i, X_{i+1}) = 0$  as well, and we shall use the convention  $0/0 = 0$ .

We now bound from above the risk of this estimator. We set

$$V_m = \left\{ \sum_{K \in m} a_K \mathbb{1}_K, \forall K \in m, a_K \in [0, +\infty) \right\}$$

and prove the following.

**Proposition 1.** *For all finite partition  $m$  of  $[0, 1]^{2d}$ ,*

$$C\mathbb{E} [H^2(s\mathbb{1}_A, \hat{s}_m)] \leq \mathbb{E} [H^2(s\mathbb{1}_A, V_m)] + \frac{1 + \log n}{n} |m|$$

where  $C = 1/(4 + \log 2)$ .

Up to a constant, the risk of  $\hat{s}_m$  is bounded by a sum of two terms. The first one corresponds to the approximation term whereas the second one corresponds to the estimation term.

An analogue upper bound on the empirical quadratic risk of this estimator may be found in Chapter 4 of Akakpo (2009). Her bound requires several assumptions on the partition  $m$  and the Markov chain although the present one requires none. However, unlike hers, we lose a logarithmic term.

**2.2. Definition of the partitions.** In this section, we shall deal with special choice of partitions  $m$ . More precisely, we consider the family of partitions defined by using the recursive algorithm developed in DeVore and Yu (1990). For  $j \in \mathbb{N}$ , we consider the set

$$\mathcal{L}_j = \left\{ \mathbf{l} = (l_1, \dots, l_{2d}) \in \mathbb{N}^{2d}, 1 \leq l_i \leq 2^j \text{ for } 1 \leq i \leq 2d \right\}$$

and define for all  $\mathbf{l} = (l_1, \dots, l_{2d}) \in \mathcal{L}_j$ ,

$$\forall i \in \{1, \dots, 2d\}, \quad I_j(l_i) = \begin{cases} \left[ \frac{l_i-1}{2^j}, \frac{l_i}{2^j} \right) & \text{if } l_i < 2^j \\ \left[ \frac{l_i-1}{2^j}, 1 \right] & \text{if } l_i = 2^j. \end{cases}$$

We then introduce the cube  $K_{j,\mathbf{l}} = \prod_{i=1}^{2d} I_j(l_i)$  and set  $\mathcal{K}_j = \{K_{j,\mathbf{l}}, \mathbf{l} \in \mathcal{L}_j\}$ .

The algorithm starts with  $[0, 1]^{2d}$ . At each step, it gets a partition of  $[0, 1]^{2d}$  into a finite family of disjoint cubes of the form  $K_{j,\mathbf{l}}$ . For any such cube, one decides to divide it into the  $4^d$  elements of  $\mathcal{K}_{j+1}$  which are contained in it, or not. The set of all such partitions that can be constructed in less than  $\ell$  steps is denoted by  $\mathcal{M}_\ell$ . We set  $\mathcal{M}_\infty = \cup_{\ell \geq 1} \mathcal{M}_\ell$ . Two examples of partitions are illustrated in Figure 1 (for  $d = 1$ ).

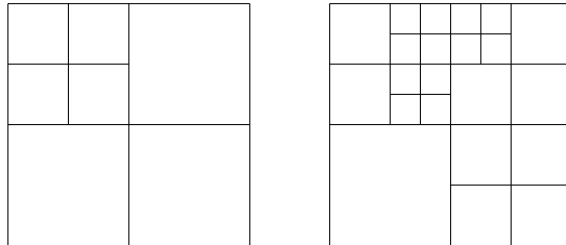


FIGURE 1. Left: example of a partition of  $\mathcal{M}_2$ . Right: example of a partition of  $\mathcal{M}_3$ .

**2.3. The selection rule.** Given  $\ell \in \mathbb{N}^* \cup \{\infty\}$ , the aim of this section is to select an estimator among the family  $\{\hat{s}_m, m \in \mathcal{M}_\ell\}$ .

For any  $K \in \cup_{m \in \mathcal{M}_\ell} m$  and any partition  $m' \in \mathcal{M}_\ell$ , let  $m' \vee K$  be the partition of  $K$  defined by

$$m' \vee K = \{K' \cap K, K' \in m', K \cap K' \neq \emptyset\}.$$

Let  $L$  be a positive number and  $\text{pen}$  be the non-negative map defined by

$$\text{pen}(m' \vee K) = L \frac{|m' \vee K| \log n}{n} \quad \text{for all } m' \in \mathcal{M}_\ell \text{ and } K \in \cup_{m \in \mathcal{M}_\ell} m.$$

Let us set  $\alpha = (1 - 1/\sqrt{2})/2$  and for all  $f, f' \in \mathbb{L}_+^1(\mathbb{X}^2, M)$ ,

$$\begin{aligned} (2) \quad T(f, f') &= \frac{1}{2n\sqrt{2}} \sum_{i=0}^{n-1} \int_{\mathbb{X}} \sqrt{f(X_i, y) + f'(X_i, y)} \left( \sqrt{f'(X_i, y)} - \sqrt{f(X_i, y)} \right) d\mu(y) \\ &\quad + \frac{1}{n\sqrt{2}} \sum_{i=0}^{n-1} \frac{\sqrt{f'(X_i, X_{i+1})} - \sqrt{f(X_i, X_{i+1})}}{\sqrt{f(X_i, X_{i+1}) + f'(X_i, X_{i+1})}} \\ &\quad + \frac{1}{2n} \sum_{i=0}^{n-1} \int_{\mathbb{X}} (f(X_i, y) - f'(X_i, y)) d\mu(y). \end{aligned}$$

We define  $\gamma$  for  $m \in \mathcal{M}_\ell$  by

$$\gamma(m) = \left\{ \sum_{K \in m} \sup_{m' \in \mathcal{M}_\ell} \left[ \left( \sum_{K' \in m'} (\alpha H^2(\hat{s}_m \mathbb{1}_{K \cap K'}, \hat{s}_{m'} \mathbb{1}_{K \cap K'}) + T(\hat{s}_m \mathbb{1}_{K \cap K'}, \hat{s}_{m'} \mathbb{1}_{K \cap K'})) \right) - \text{pen}(m' \vee K) \right] \right\} + 2\text{pen}(m).$$

Finally, we select  $\hat{m}$  among  $\mathcal{M}_\ell$  as any partition satisfying

$$(3) \quad \gamma(\hat{m}) \leq \inf_{m \in \mathcal{M}_\ell} \gamma(m) + \frac{1}{n}$$

and consider the resulting estimator  $\hat{s} = \hat{s}_{\hat{m}}$ .

**Remarks.** The estimator  $\hat{s} = \hat{s}(L, \ell)$  depends on the choices of two quantities  $L > 0, \ell \in \mathbb{N}^* \cup \{\infty\}$ . We shall see in the next section that  $L$  can be chosen as an universal numerical constant. As to  $\ell$ , from a theoretical point of view, it can be chosen as  $\ell = \infty$ . In practice, we recommend to take it as large as possible. Nevertheless, the larger  $\ell$ , the longer it takes to compute the estimator. A practical algorithm in view of computing  $\hat{m}$  will be detailed in the appendix.

The selection procedure we use may look somewhat unusual. It can be seen as a mix between a procedure based on a contrast function (which is usually easy to implement) and a procedure based on a robust test (the functional  $T(f, f')$ , which can be seen as a robust test between  $f, f'$ , will allow us to obtain risk bounds with respect to a Hellinger-type distance). This functional is inspired from the variational formula for the Hellinger affinity described in Section 2 of Baraud (2010).

**2.4. An oracle inequality.** The main result of this section is the following.

**Theorem 2.** *There exists an universal constant  $L_0 > 0$  such that, for all  $L \geq L_0$ ,  $\ell \in \mathbb{N}^* \cup \{\infty\}$ , the estimator  $\hat{s} = \hat{s}(L, \ell)$  satisfies*

$$(4) \quad C\mathbb{E} [H^2(s\mathbb{1}_A, \hat{s})] \leq \inf_{m \in \mathcal{M}_\ell} \{ \mathbb{E} [H^2(s\mathbb{1}_A, V_m)] + \text{pen}(m) \}$$

where  $C$  is an universal positive constant.

In the literature, oracle inequalities with a random quadratic loss for piecewise constant estimators have been obtained in Lacour (2007) and Akakpo and Lacour (2011). Their procedures require a priori assumptions on the transition density and the Markov chain although ours requires none (except homogeneity). However, unlike theirs, our risk bound involves an extra logarithmic term. We do not know whether this term is necessary or not.

In the proof, we obtain an upper bound for  $L_0$  which is unfortunately very rough and useless in practice. It seems difficult to obtain a sharp bound on  $L_0$  from the theory and we have rather carried out a simulation study in order to tune  $L_0$  (see Section 3).

**2.5. Risk bounds with respect to a deterministic loss.** Although the distance  $H$  is natural, we are interested in controlling the risk associated to a deterministic distance. To do so, we shall make a posteriori assumptions on the Markov chain.

**Assumption 1.** *The sequence  $(X_i)_{i \geq 0}$  is stationary and admits a stationary density  $\varphi$  with respect to the Lebesgue measure  $\mu$  on  $\mathbb{R}^d$ . There exists  $\kappa_0 > 0$  such that  $\varphi(x) \geq \kappa_0$  for all  $x \in [0, 1]^d$ .*

We introduce  $\mathbb{L}_+^1([0, 1]^{2d}, (\varphi \cdot \mu) \otimes \mu)$  the cone of integrable and non-negative functions on  $[0, 1]^{2d}$  with respect to the product measure  $(\varphi \cdot \mu) \otimes \mu$ . We endow  $\mathbb{L}_+^1([0, 1]^{2d}, (\varphi \cdot \mu) \otimes \mu)$  with the distance  $h$  defined by

$$\forall f, f' \in \mathbb{L}_+^1([0, 1]^{2d}, (\varphi \cdot \mu) \otimes \mu), \quad h^2(f, f') = \frac{1}{2} \int_{[0, 1]^{2d}} \left( \sqrt{f(x, y)} - \sqrt{f'(x, y)} \right)^2 \varphi(x) dx dy.$$

In our results, we shall need the  $\beta$ -mixing properties of the Markov chain. We set for all  $q \in \mathbb{N}^*$

$$\beta_q = \int_{\mathbb{R}^{2d}} |s^{(q)}(x, y) - \varphi(y)| \varphi(x) dx dy$$

where  $s^{(q)}(x, \cdot)$  is the density of the conditional law  $\mathcal{L}(X_q | X_0 = x)$  with respect to the Lebesgue measure. We refer to Doukhan (1994) and Bradley (2005) for more details on the  $\beta$ -mixing coefficients.

**Theorem 3.** *Under Assumption 1, the estimator  $\hat{s}$  built in Section 2.3 with  $\ell \in \mathbb{N}^*$  and  $L \geq L_0$ , satisfies*

$$C\mathbb{E} [h^2(s\mathbb{1}_A, \hat{s})] \leq \inf_{m \in \mathcal{M}_\ell} \{ h^2(s\mathbb{1}_A, V_m) + \text{pen}(m) \} + \frac{R_n(\ell)}{n}$$

where

$$(5) \quad R_n(\ell) = n 2^{3\ell d} \inf_{1 \leq q \leq n} \left\{ \exp \left( -\frac{\kappa_0}{10} \frac{n}{q 2^{\ell d}} \right) + n \beta_q / q \right\}$$

and where  $C$  is an universal positive constant.

This result is interesting when the remainder term  $R_n(\ell)/n$  is small enough, that is when  $2^{\ell d}$  is small compared to  $n$  and when the sequence  $(\beta_q)_{q \geq 1}$  goes to 0 fast enough. More precisely,  $R_n(\ell)$  can be bounded independently of  $n, \ell$  whenever  $\ell, d, n$  and the  $\beta_q$  coefficients satisfy the following.

- If the chain is geometrically  $\beta$ -mixing, that is if there exists  $b_1 > 0$  such that  $\beta_q \leq e^{-b_1 q}$ , then

$$R_n(\ell) \leq n^2 2^{3\ell d + 1} \left[ \exp(-b_1 n) + \exp\left(-\frac{\kappa_0}{10} \frac{n}{2^{\ell d}}\right) + \exp\left(-\sqrt{\frac{\kappa_0 b_1}{40}} \frac{n}{2^{\ell d}}\right) \right].$$

In particular, if  $\ell, d, n$  are such that  $2^{\ell d} \leq n / \log^3 n$ ,  $R_n(\ell)$  is upper bounded by a constant depending only on  $\kappa_0, b_1$ .

- If the chain is arithmetically  $\beta$ -mixing, that is if there exists  $b_2 > 0$  such that  $\beta_q \leq q^{-b_2}$ , then

$$R_n(\ell) \leq \frac{C'(b_2)}{\kappa_0^{b_2+1}} \frac{2^{(4+b_2)\ell d} \log^{b_2+1} \left(1 + \frac{\kappa_0 n}{2^{\ell d}}\right)}{n^{b_2-1}}$$

where  $C'(b_2)$  depends only on  $b_2$ . Consequently, if  $2^{\ell d} \leq n^{1-\zeta} / \log n$  and  $b_2 \geq 5/\zeta - 4$  for  $\zeta \in (0, 1)$ ,  $R_n(\ell)$  is upper bounded by a constant depending only on  $\kappa_0, b_2$ .

**2.6. Rates of convergence.** The aim of this section is to obtain uniform risk bounds over classes of smooth transition densities for our estimator.

**2.6.1. Hölder spaces.** Given  $\sigma \in (0, 1]$ , we say that a function  $f$  belongs to the Hölder space  $\mathcal{H}^\sigma([0, 1]^{2d})$  if there exists  $|f|_\sigma \in \mathbb{R}_+$  such that for all  $(x_1, \dots, x_{2d}) \in [0, 1]^{2d}$  and all  $1 \leq j \leq 2d$ , the functions  $f_j(\cdot) = f(x_1, \dots, x_{j-1}, \cdot, x_{j+1}, \dots, x_{2d})$  satisfy

$$|f_j(x) - f_j(y)| \leq |f|_\sigma |x - y|^\sigma \quad \text{for all } x, y \in [0, 1].$$

When the restriction of  $\sqrt{s}$  to  $A = [0, 1]^{2d}$  is Hölderian, we deduce from (4) the following.

**Corollary 1.** *For all  $\sigma \in (0, 1]$  and  $\sqrt{s}|_A \in \mathcal{H}^\sigma([0, 1]^{2d})$ , the estimator  $\hat{s} = \hat{s}(L_0, \infty)$  satisfies*

$$C\mathbb{E} [H^2(s\mathbb{1}_A, \hat{s})] \leq (d|\sqrt{s}|_A|_\sigma)^{\frac{2d}{d+\sigma}} \left( \frac{\log n}{n} \right)^{\frac{\sigma}{\sigma+d}} + \frac{\log n}{n}$$

where  $C$  is an universal positive constant.

**2.6.2. Besov spaces.** A thinner way to measure the smoothness of the transition density is to assume that  $\sqrt{s}|_A$  belongs to a Besov space. We refer to Section 3 of DeVore and Yu (1990) for a definition of this space. We say that the Besov space  $\mathcal{B}_q^\sigma(\mathbb{L}^p([0, 1]^{2d}))$  is homogeneous when  $p \geq 2$  and inhomogeneous otherwise. We set for all  $p \in (1, +\infty)$  and  $\sigma \in (0, 1)$ ,

$$\mathcal{B}^\sigma(\mathbb{L}^p([0, 1]^{2d})) = \begin{cases} \mathcal{B}_p^\sigma(\mathbb{L}^p([0, 1]^{2d})) & \text{if } p \in (1, 2) \\ \mathcal{B}_\infty^\sigma(\mathbb{L}^p([0, 1]^{2d})) & \text{if } p \in [2, +\infty), \end{cases}$$

and denote by  $|\cdot|_{p,\sigma}$  the semi norm of  $\mathcal{B}^\sigma(\mathbb{L}^p([0, 1]^{2d}))$ . We make the following assumption to deduce from (4) risk bounds over these spaces.

**Assumption 2.** *There exists  $\kappa > 0$  such that for all  $i \in \{0, \dots, n-1\}$ ,  $X_i$  admits a density  $\varphi_i$  with respect to the Lebesgue measure  $\mu$  such that  $\varphi_i(x) \leq \kappa$  for all  $x \in [0, 1]^d$ .*



Note that we do not require that the chain be either stationary or mixing.

Let  $(\mathbb{L}^2([0, 1]^{2d}, \mu \otimes \mu), d_2)$ , be the metric space of square integrable functions on  $[0, 1]^{2d}$  with respect to the Lebesgue measure. Under the above assumption, we deduce from (4) that

$$C\mathbb{E} [H^2(s\mathbb{1}_A, \hat{s}_m)] \leq \inf_{m \in \mathcal{M}_\ell} \left\{ \kappa d_2^2(\sqrt{s}|_A, V_m) + L_0 \frac{|m| \log n}{n} \right\}.$$

When  $\sqrt{s}|_A$  belongs to a Besov space, the right-hand side of this inequality can be upper bounded thanks to the approximation theorems of DeVore and Yu (1990).

**Corollary 2.** *Suppose that Assumption 2 holds. For all  $p \in (2d/(d+1), +\infty)$ ,  $\sigma \in (2d(1/p - 1/2)_+, 1)$  and  $\sqrt{s}|_A \in \mathcal{B}^\sigma(\mathbb{L}^p([0, 1]^{2d}))$ , the estimator  $\hat{s} = \hat{s}(L_0, \infty)$  satisfies*

$$(6) \quad C'\mathbb{E} [H^2(s\mathbb{1}_A, \hat{s})] \leq |\sqrt{s}|_A|_{p, \sigma}^{\frac{2d}{d+\sigma}} \left( \frac{\log n}{n} \right)^{\frac{\sigma}{\sigma+d}} + \frac{\log n}{n}$$

where  $C' > 0$  depends only on  $\kappa, \sigma, d, p$ .

More precisely, it is shown in the proof that the estimators  $\hat{s} = \hat{s}(L_0, \ell)$  satisfy (6) when  $\ell$  is large enough (when  $\ell \geq d^{-1}(\log 2)^{-1} \log n$ ).

Rates of convergence for the deterministic loss  $h$  can be established by using Theorem 3 instead of Theorem 2. For instance, if the chain is geometrically  $\beta$ -mixing, we may choose  $\ell$  the smallest integer larger than  $d^{-1}(\log 2)^{-1} \log(n/\log^3 n)$ , in which case the estimator  $\hat{s} = \hat{s}(L_0, \ell)$  achieves the rate  $(\log n/n)^{\sigma/(\sigma+d)}$  over the Besov spaces  $\mathcal{B}^\sigma(\mathbb{L}^p([0, 1]^{2d}))$ ,  $p \in (2d/(d+1), +\infty)$ ,  $\sigma \in (\sigma_1(p, d), 1)$  where

$$\sigma_1(p, d) = \frac{d}{4} \left( -1 + 4(1/p - 1/2)_+ + \sqrt{1 + 24(1/p - 1/2)_+ + 16(1/p - 1/2)_+^2} \right).$$

If the chain is arithmetically  $\beta$ -mixing with  $b_q \leq q^{-6}$ , choosing  $\ell$  the smallest integer larger than  $d^{-1}(2 \log 2)^{-1} \log(n/\log n)$  allows us to recover the same rate of convergence when  $\sigma \in (\sigma_2(p, d), 1)$  where

$$\sigma_2(p, d) = d \left( (1/p - 1/2)_+ + \sqrt{2(1/p - 1/2)_+ + (1/p - 1/2)_+^2} \right).$$

We refer the reader to Section 6.7 for a proof of these two results.

In the literature, Lacour (2007) obtained a rate of order  $n^{-\sigma/(\sigma+1)}$  over  $\mathcal{B}^\sigma(\mathbb{L}^2([0, 1]^2))$ , which is slightly faster but her approach prevents her to deal with inhomogeneous Besov spaces and requires the prior knowledge of a suitable upper bound on the supremum norm of  $s$ . As far as we know, the rates that have been established in the other papers hold only when  $\sigma > 1$ .

### 3. SIMULATIONS.

In this section, we present a simulation study to evaluate the performance of our estimator in practice. We shall simulate several Markov chains and estimate their transition densities by using our procedure.

**3.1. Examples of Markov chains.** We consider Markov chains of the form

$$X_{k+1} = F(X_k, U_k)$$

where  $F$  is some known function and where  $U_k$  is a random variable independent of  $(X_0, \dots, X_k)$ .

For the sake of comparison, we begin to deal with examples that have already been considered in the simulation study of Akakpo and Lacour (2011). In each of these examples,  $U_k$  is a standard Gaussian random variable.

**Example 1.**  $X_{k+1} = 0.5X_k + (1 + U_k)/4$

**Example 2.**  $X_{k+1} = 12^{-1} (6 + \sin(12X_k - 6) + (\cos(X_k - 6) + 3)U_k)$

**Example 3.**

$$X_{k+1} = \frac{1}{3} (X_k + 1) + \left( \frac{1}{9} - \frac{1}{23} \left( \frac{1}{2} \beta(5X_k/3, 4, 4) + \frac{1}{20} \beta((5X_k - 2)/3, 400, 400) \right) \right) U_k$$

where  $\beta(\cdot, a, b)$  is the density of the  $\beta$  distribution with parameters  $a$  and  $b$ .

**Example 4.**

$$X_{k+1} = \frac{1}{4} (g(X_k) + 1) + \frac{1}{8} U_k$$

where  $g$  is defined by

$$g(x) = \frac{9\sqrt{2}}{4\sqrt{\pi}} \exp(-18(x - 1/2)^2) + \frac{9\sqrt{2}}{4\sqrt{\pi}} \exp(-162(x - 3/4)^2) \quad \text{for all } x \in \mathbb{R}.$$

At first sight, Examples 1 and 2 may seem to be different than those of Akakpo and Lacour (2011). Actually, we just have rescaled the data in order to estimate on  $[0, 1]^2$ . The statistical problem is the same. According to Akakpo and Lacour (2011), we set  $p$  large ( $p = 10^4$ ) and we estimate the transition densities of Examples 1, 2, 3 and 4 from  $(X_p, \dots, X_{n+p})$  so that the chain is approximatively stationary.

We also propose to consider the following examples. In Example 5,  $U_k$  is a centred Gaussian random variable with variance  $1/2$ , in Example 6,  $U_k$  admits the density

$$f(x) = \frac{5\sqrt{2}}{2\sqrt{\pi}} [\exp(-50(x - 1)^2) + \exp(-50x^2)]$$

with respect to the Lebesgue measure, and in Example 7,  $U_k$  is an exponential random variable with parameter 1.

**Example 5.**  $X_{k+1} = 0.5X_k + (1 + U_k)/4$ .

**Example 6.**  $X_{k+1} = 0.5(X_k + U_k)$ .

**Example 7.**  $X_{k+1} = X_k/(50X_k + 1) + X_k U_k$ .

We set  $X_0 = 1/2$  and estimate  $s$  from  $(X_0, \dots, X_n)$ . These last three Markov chains are not stationary. Their transition densities are rather isotropic and inhomogeneous. The transition density of Example 7 is unbounded.

In what follows, our selection rule will always be applied with  $L = 0.03$  (whatever,  $\ell$ ,  $n$  and the Markov chain).

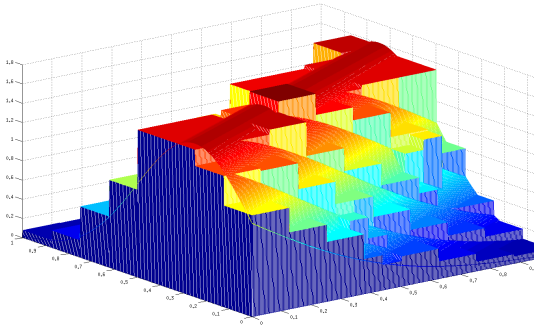
**3.2. Choice of  $\ell$ .** We discuss the choice of  $\ell$  by simulating the preceding examples with  $n = 10^3$  and by applying our selection rule for each value of  $\ell \in \{1, \dots, 10\}$ . The results are summarized below.

$\ell$	Ex 1	Ex 2	Ex 3	Ex 4	Ex 5	Ex 6	Ex 7
1	0.031	0.046	0.299	0.181	0.089	0.291	0.358
2	0.011	0.015	0.087	0.107	0.024	0.170	0.241
3	0.011	0.014	0.026	0.058	0.013	0.067	0.156
4	0.011	0.018	0.026	0.035	0.015	0.046	0.113
5	0.011	0.018	0.022	0.038	0.015	0.048	0.098
6	0.011	0.018	0.022	0.038	0.015	0.048	0.065
7	0.011	0.018	0.024	0.038	0.015	0.048	0.044
8	0.011	0.018	0.024	0.038	0.015	0.048	0.040
9	0.011	0.018	0.024	0.038	0.015	0.048	0.040
10	0.011	0.018	0.024	0.038	0.015	0.048	0.040

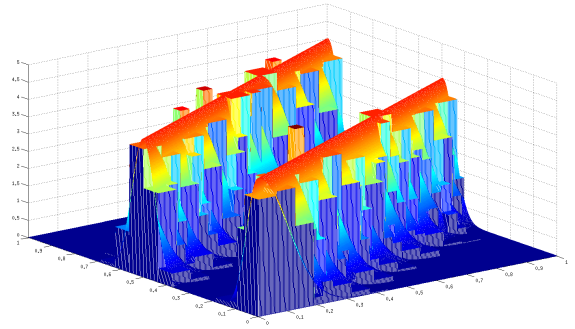
FIGURE 2. Hellinger risk  $H^2(s\mathbb{1}_{[0,1]^2}, \hat{s})$ .

When  $\ell$  grows up, the risk of our estimator tends to decrease and then stabilize. The best choice of  $\ell$  is obviously unknown in practice but this array shows that a good way for choosing  $\ell$  is to take it as large as possible. This is theoretically justified by Theorem 2 since the right-hand side of inequality (4) is a non-increasing function of  $\ell$ .

**3.3. An illustration.** We apply our procedure for Examples 1 and 6 with  $n = 10^4$ ,  $\ell = 7$ . We get two estimators and draw them with the corresponding transition density in Figure 3.



Example 1.



Example 6.

FIGURE 3. Estimator and transition density.

This shows that the selected partition is thinner (respectively wider) to the points where the transition density is changing rapidly (respectively slower), and is thus rather well adapted to the target function  $s$ .

**3.4. Comparison with other procedures.** In this section, we compare our selection rule with the oracle estimator and with the piecewise constant estimator of Akakpo and Lacour (2011).

The procedure of Akakpo and Lacour (2011) amounts to selecting an estimator among  $\{\hat{s}_m, m \in \mathcal{M}'\}$  where  $\hat{s}_m$  is defined by (1) and where  $\mathcal{M}'$  is a collection of irregular partitions on  $[0, 1]^2$ . Precisely, with their notations, we apply it with  $J_\star = 5$ ,  $\text{pen}(m) = 3\|s_{1A}\|_\infty |m|/n$  and with  $\text{pen}(m) = 3\|\hat{s}_{m^\bullet}\|_\infty |m|/n$  where  $m^\bullet$  is a partition suitably chosen (following the recommendations of Akakpo and Lacour (2011), that is  $J_\bullet = 3$ ). These two estimators are denoted by  $\hat{s}^{(1)}$  and  $\hat{s}^{(2)}$  respectively. Notice that these penalties, which are used in their simulation study, are not the ones prescribed by their theory. Their theoretical penalties also depend on a positive lower bound on the stationary density.

We denote by  $\hat{s}^{(0)}$  the oracle estimator, that is the estimator defined as being a minimizer of the map  $m \mapsto H^2(s\mathbb{1}_{[0,1]^2}, \hat{s}_m)$  for  $m \in \mathcal{M}_7$ . This estimator is the best estimator of the family  $\{\hat{s}_m, m \in \mathcal{M}_7\}$  and is known since the data are simulated. We consider the random variables

$$\mathcal{R}_i = \frac{H^2(s\mathbb{1}_{[0,1]^2}, \hat{s})}{H^2(s\mathbb{1}_{[0,1]^2}, \hat{s}^{(i)})} \quad \text{for } i = 0, 1, 2$$

and denote by  $q_0(\alpha)$  the  $\alpha$ -quantile of  $\mathcal{R}_0$ . Results obtained are given in Figure 4.

	Ex 1	Ex 2	Ex 3	Ex 4	Ex 5	Ex 6	Ex 7
$\mathbb{E}[H^2(s\mathbb{1}_{[0,1]^2}, \hat{s})]$	0.011	0.017	0.022	0.038	0.018	0.052	0.049
$\mathbb{E}[H^2(s\mathbb{1}_{[0,1]^2}, \hat{s}^{(0)})]$	0.007	0.011	0.015	0.028	0.012	0.037	0.041
$q_0(0.5)$	1.473	1.513	1.443	1.369	1.422	1.420	1.200
$q_0(0.75)$	1.698	1.627	1.557	1.440	1.575	1.481	1.244
$q_0(0.9)$	1.921	1.834	1.683	1.509	1.749	1.543	1.290
$q_0(0.95)$	2.113	1.965	1.770	1.558	1.839	1.590	1.317
$\mathbb{E}[H^2(s\mathbb{1}_{[0,1]^2}, \hat{s}^{(1)})]$	0.017	0.018	0.028	0.058	0.024	0.103	-
$\mathbb{P}(\mathcal{R}_2 \leq 1)$	0.964	0.740	0.908	1	0.984	1	-
$\mathbb{E}[H^2(s\mathbb{1}_{[0,1]^2}, \hat{s}^{(2)})]$	0.013	0.018	0.028	0.062	0.023	0.096	0.133
$\mathbb{P}(\mathcal{R}_3 \leq 1)$	0.832	0.748	0.928	1	0.948	1	1

FIGURE 4. Risks for simulated data with  $n = 1000$  averaged over 250 samples.

**3.5. Comparison with a quadratic empirical risk.** In Akakpo and Lacour (2011), the risks of the estimators are evaluated with a empirical quadratic norm and we can also compare the performances of our estimator to theirs by using this risk.

To do so, let us denote by  $\|\cdot\|_n$  the empirical quadratic norm defined by

$$\|f\|_n^2 = \frac{1}{n} \sum_{i=1}^n \int_{\mathbb{R}} f^2(X_i, x) dx \quad \text{for all } f \in \mathbb{L}^2(\mathbb{R}^2, M)$$

and set for  $i \in \{1, 2\}$ ,

$$\mathcal{R}'_i = \frac{\|s\mathbb{1}_{[0,1]^2} - \hat{s}\|_n^2}{\|s\mathbb{1}_{[0,1]^2} - \hat{s}^{(i)}\|_n^2}.$$

The results obtained are presented in Figure 5. They are very similar to those of Figure 4.

	Ex 1	Ex 2	Ex 3	Ex 4	Ex 5	Ex 6	Ex 7
$\mathbb{E}[\ s\mathbb{1}_{[0,1]^2} - \hat{s}\ _n^2]$	0.064	0.108	0.229	0.319	0.116	0.528	2.82
$\mathbb{E}[\ s\mathbb{1}_{[0,1]^2} - \hat{s}^{(1)}\ _n^2]$	0.147	0.133	0.257	0.423	0.205	0.743	-
$\mathbb{P}(\mathcal{R}_2' \leq 1)$	0.980	0.820	0.788	0.984	0.992	1	-
$\mathbb{E}[\ s\mathbb{1}_{[0,1]^2} - \hat{s}^{(2)}\ _n^2]$	0.091	0.129	0.262	0.418	0.159	0.739	6.08
$\mathbb{P}(\mathcal{R}_3' \leq 1)$	0.864	0.780	0.792	0.980	0.940	1	1

FIGURE 5. Risks for simulated data with  $n = 1000$  averaged over 250 samples.

#### 4. A GENERAL PROCEDURE.

In Section 2, we used our selection rule to establish the oracle inequality (4), from which we deduced rates of convergence over Besov spaces  $\mathcal{B}^\sigma(\mathbb{L}^p([0, 1]^{2d}))$  with  $\sigma$  lower than 1. We now aim at obtaining rates for more general spaces of functions. This includes Besov spaces with regularity index larger than 1 and spaces corresponding to structural assumptions on  $s$ . We propose a second procedure to reach this goal.

The Markov chain takes its values into  $\mathbb{X}$  and we estimate  $s$  on a subset  $A$  of the form  $A = A_1 \times A_2$ . We always assume that  $n > 3$ .

**4.1. Procedure and preliminary result.** Our second procedure is defined as follows. Let  $\alpha = (1 - 1/\sqrt{2})/2$ ,  $L > 0$ ,  $S$  be an at most countable set of  $\mathbb{L}_+^1(\mathbb{X}^2, M)$  and  $\Delta_S \geq 1$  be a map on  $S$ .

We define the application  $\wp$  on  $S$  by

$$\wp(f) = \sup_{f' \in S} \left[ \alpha H^2(f, f') + T(f, f') - L \frac{\Delta_S(f')}{n} \right] + L \frac{\Delta_S(f)}{n} \quad \text{for all } f \in S.$$

We select  $\hat{s}$  among  $S$  as any element of  $S$  satisfying

$$\wp(\hat{s}) \leq \inf_{f \in S} \wp(f) + \frac{1}{n}.$$

We prove the following.

**Proposition 4.** *Suppose that  $f(x) = 0$  for all  $f \in S$  and  $x \in \mathbb{X}^2 \setminus A$  and that  $\sum_{f \in S} e^{-\Delta_S(f)} \leq 1$ . There exists an universal constant  $L_0 > 0$  such that if  $L \geq L_0$ , the estimator  $\hat{s}$  satisfies*

$$(7) \quad C\mathbb{E}[H^2(s\mathbb{1}_A, \hat{s})] \leq \mathbb{E} \left[ \inf_{f \in S} \left\{ H^2(s\mathbb{1}_A, f) + L \frac{\Delta_S(f)}{n} \right\} \right]$$

where  $C$  is an universal positive constant.

**4.2. A general model selection theorem.** We shall deduce from the above proposition a model selection theorem by choosing suitably  $S$ . To do so, we consider the following assumption.

**Assumption 3.** *For all  $i \in \{1, \dots, n-1\}$ ,  $X_i$  admits a density  $\varphi_i$  with respect to some known measure  $\nu$  such that  $\nu(A_1) = 1$ . Moreover, there exists  $\kappa$  such that  $\varphi_i(x) \leq \kappa$  for all  $x \in A_1$  and  $i \in \{1, \dots, n-1\}$ .*

We define  $\mathbb{L}^2(A, \nu \otimes \mu)$  the space of square integrable functions on  $A$  with respect to the product measure  $\nu \otimes \mu$ , and we endow it with its natural distance

$$d^2(f, f') = \int_A (f(x, y) - f'(x, y))^2 d\nu(x) d\mu(y) \quad \text{for all } f, f' \in \mathbb{L}^2(A, \nu \otimes \mu).$$

Hereafter, a model  $V$  is a (non-trivial) finite dimensional linear space of  $\mathbb{L}^2(A, \nu \otimes \mu)$ .

Let us explain how to obtain a model selection theorem when Assumption 3 holds. Let  $\mathbb{V}$  be a collection of models  $V$  and let  $(\Delta(V))_{V \in \mathbb{V}}$  be a family of non-negative numbers such that  $\sum_{V \in \mathbb{V}} e^{-\Delta(V)} \leq 1$ . For each model  $V \in \mathbb{V}$ , we consider an orthonormal basis  $(f_1, \dots, f_{\dim V})$  of  $V$  and set

$$T_V = \left\{ \sum_{i=1}^{\dim V} \alpha_i f_i, \alpha_i \in \frac{2}{\sqrt{n \dim V}} \mathbb{Z} \right\}.$$

We deduce from Lemma 5 of Birgé (2006) that the cardinal of  $S_V = \{f_+^2 \mathbb{1}_A, f \in T_V, d(f, 0) \leq 2\}$  is upper bounded by  $|S_V| \leq (30n)^{\dim V/2}$ . We then use the above procedure with  $S = \cup_{V \in \mathbb{V}} S_V$  and

$$\Delta_S(f) = \inf_{\substack{V \in \mathbb{V} \\ S_V \ni f}} \{ \Delta(V) + (\dim V) \log(30n)/2 \} \quad \text{for all } f \in S.$$

This yields to an estimator  $\hat{s}$  such that

$$C' \mathbb{E} [H^2(s \mathbb{1}_A, \hat{s})] \leq \inf_{V \in \mathbb{V}} \left\{ \kappa \left( \inf_{\substack{f \in T_V \\ d(f, 0) \leq 2}} d^2(\sqrt{s}|_A, f) \right) + \frac{\Delta(V) + \dim(V) \log n}{n} \right\}$$

where  $C'$  is an universal positive constant. Since  $d(\sqrt{s}|_A, 0) \leq 1$ ,

$$\inf_{\substack{f \in T_V \\ d(f, 0) \leq 2}} d^2(\sqrt{s}|_A, f) = d^2(\sqrt{s}|_A, T_V).$$

For all  $f' \in V$ , there exists  $f \in T_V$  such that  $d^2(f, f') \leq n^{-1}$  and thus

$$d^2(\sqrt{s}|_A, T_V) \leq 2d^2(\sqrt{s}|_A, V) + \frac{2}{n}.$$

Precisely, we have proved:

**Theorem 5.** *Suppose that Assumption 3 holds. Let  $\mathbb{V}$  be an at most countable collection of models. Let  $(\Delta(V))_{V \in \mathbb{V}}$  be a family of non-negative numbers such that*

$$\sum_{V \in \mathbb{V}} e^{-\Delta(V)} \leq 1.$$

*There exists an estimator  $\hat{s}$  such that*

$$C \mathbb{E} [H^2(s \mathbb{1}_A, \hat{s})] \leq \inf_{V \in \mathbb{V}} \left\{ d^2(\sqrt{s}|_A, V) + \frac{\Delta(V) + \dim(V) \log n}{n} \right\}$$

where  $C > 0$  depends only on  $\kappa$ .

The condition  $\sum_{\mathbb{V} \in \mathbb{V}} e^{-\Delta(V)} \leq 1$  can be interpreted as a (sub)probability on the collection  $\mathbb{V}$ . The more complex the family  $\mathbb{V}$ , the larger the weights  $\Delta(V)$ . When one can choose  $\Delta(V)$  of order  $\dim(V)$ , which means that the family  $\mathbb{V}$  of models does not contains too many models per dimension, the estimator  $\hat{s}$  achieves the best trade-off (up to a constant) between the approximation and the variance terms.

This theorem holds under an assumption that is very mild and weaker than those of Lacour (2007), Akakpo and Lacour (2011) and Cl  mencon (2000). Birg   (2012) proved a general oracle inequality when there exist integers  $k \geq 1$  and  $l \geq 0$  and positive numbers  $\rho, \varrho$  such that

$$\varrho \leq \frac{1}{k} \sum_{j=1}^k s^{(l+j)}(x, y) \leq \rho \quad \text{for all } x, y \in \mathbb{X}$$

where the parameters  $k, l, \varrho$  are known. Our assumption is then satisfied for the Markov chain  $(X_{l+1}, \dots, X_n)$  with  $\nu = \mu$  and  $\kappa = k\rho$ .

We shall consider subsets  $\mathcal{F} \subset \mathbb{L}^2(A, \nu \otimes \mu)$  corresponding to smoothness or structural assumptions on  $\sqrt{s}|_A$ . For such an  $\mathcal{F}$ , we associate a collection  $\mathbb{V}$  and deduce from Theorem 5 a risk bound for the estimator  $\hat{s}$  when  $\sqrt{s}|_A$  belongs to  $\mathcal{F}$ . This set is a generic notation and will change from section to section. In the remaining part of this paper, we shall always choose  $\mathbb{X}^2 = \mathbb{R}^{2d}$ ,  $A = [0, 1]^{2d}$  and  $\mu$  the Lebesgue measure.

**4.3. Smoothness assumptions.** We have introduced in Section 2.6 the isotropic Besov spaces  $\mathcal{B}_q^\sigma(\mathbb{L}^p([0, 1]^{2d}))$  where  $\sigma \in (0, 1)$ . In this section, we consider the anisotropic Besov spaces  $\mathcal{B}_q^\sigma(\mathbb{L}^p([0, 1]^{2d}))$  where  $\sigma = (\sigma_1, \dots, \sigma_{2d})$  belongs to  $(0, +\infty)^{2d}$ .

Intuitively, a function  $f$  on  $[0, 1]^{2d}$  belongs to  $\mathcal{B}_q^\sigma(\mathbb{L}^p([0, 1]^{2d}))$  if, for all  $j \in \{1, \dots, 2d\}$ , and  $x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_{2d} \in [0, 1]$  the function

$$x_j \mapsto f(x_1, \dots, x_{j-1}, x_j, x_{j+1}, \dots, x_{2d})$$

belongs to  $\mathcal{B}_q^{\sigma_j}(\mathbb{L}^p([0, 1]))$ . In particular, for all  $\sigma \in (0, +\infty)$ ,

$$\mathcal{B}_q^\sigma(\mathbb{L}^p([0, 1]^{2d})) = \mathcal{B}_q^{(\sigma, \dots, \sigma)}(\mathbb{L}^p([0, 1]^{2d})).$$

A definition of the anisotropic Besov spaces may be found in Hochmuth (2002) (for  $d = 1$ ) and in Akakpo (2009) (for larger values of  $d$ ). We also consider the space  $\mathcal{H}^\sigma([0, 1]^{2d})$  of anisotropic H  lderian functions on  $[0, 1]^{2d}$  with regularity  $\sigma$ . A precise definition of this space may be found in Section 3.1.1 of Baraud and Birg   (2011) (among other references).

For all  $\sigma = (\sigma_1, \dots, \sigma_{2d}) \in (0, +\infty)^{2d}$ , we denote by  $\bar{\sigma}$  the harmonic mean of  $\sigma$ :

$$\frac{1}{\bar{\sigma}} = \frac{1}{2d} \sum_{i=1}^{2d} \frac{1}{\sigma_i}.$$

We set for all  $p \in (0, +\infty]$ ,

$$\mathcal{B}^\sigma(\mathbb{L}^p([0, 1]^{2d})) = \begin{cases} \mathcal{B}_\infty^\sigma(\mathbb{L}^p([0, 1]^{2d})) & \text{if } p \in (0, 1] \\ \mathcal{B}_p^\sigma(\mathbb{L}^p([0, 1]^{2d})) & \text{if } p \in (1, 2) \\ \mathcal{B}_\infty^\sigma(\mathbb{L}^p([0, 1]^{2d})) & \text{if } p \in [2, +\infty) \\ \mathcal{H}^\sigma([0, 1]^{2d}) & \text{if } p = \infty \end{cases}$$

and denote by  $|\cdot|_{p,\sigma}$  the semi norm associated to the space  $\mathcal{B}^\sigma(\mathbb{L}^p([0, 1]^{2d}))$ .

In this section, we are interesting in obtaining a bound risk when  $\sqrt{s}|_A$  belongs to the space

$$\mathcal{B}([0, 1]^{2d}) = \bigcup_{p \in (0, +\infty]} \left( \bigcup_{\substack{\sigma \in (0, +\infty)^d \\ \bar{\sigma} > 2d(1/p - 1/2)_+}} \mathcal{B}^\sigma(\mathbb{L}^p([0, 1]^{2d})) \right).$$

Families of linear spaces possessing good approximation properties with respect to the elements of  $\mathcal{F} = \mathcal{B}([0, 1]^{2d})$  can be found in Theorem 1 of Akakpo (2012). We then deduce from Theorem 5,

**Corollary 3.** *Suppose that Assumption 3 holds with  $\mathbb{X} = \mathbb{R}^d$ ,  $A = [0, 1]^{2d}$  and with  $\nu \otimes \mu$  the Lebesgue measure. There exists an estimator  $\hat{s}$  such that for all  $\sqrt{s}|_A \in \mathcal{B}([0, 1]^{2d})$ ,*

$$C\mathbb{E} [H^2(s\mathbb{1}_A, \hat{s})] \leq |\sqrt{s}|_A|_{p,\sigma}^{2d/(d+\bar{\sigma})} \left( \frac{\log n}{n} \right)^{\bar{\sigma}/(\bar{\sigma}+d)} + \frac{\log n}{n}$$

where  $p \in (0, +\infty]$ ,  $\sigma \in (0, +\infty)^{2d}$ ,  $\bar{\sigma} > 2d(1/p - 1/2)_+$  are such that  $\sqrt{s}|_A \in \mathcal{B}^\sigma(\mathbb{L}^p([0, 1]^{2d}))$  and where  $C > 0$  depends only on  $\kappa, d, p, \sigma$ .

To our knowledge, the only statistical procedures that can adapt both to possible inhomogeneity and anisotropy of  $s$  are those of Akakpo and Lacour (2011) and Birgé (2012). The losses are different, but the rates are the same as ours (up to the logarithmic term). In view of our assumptions, we do not know if the logarithmic term can be avoided.

In the following sections, we consider classes  $\mathcal{F}$  corresponding to structural assumptions on  $\sqrt{s}|_A$ . More precisely, rates of convergence when the chain is autoregressive with constant conditional variance (respectively non constant conditional variance) are established in Section 4.4 (respectively Section 4.5).

**4.4. AR model.** In this section, we assume that  $X_{n+1} = g(X_n) + \varepsilon_n$  where  $g$  is an unknown function and where the  $\varepsilon_n$ 's are unobserved identically distributed random variables. Many papers are devoted to the estimation of the regression function  $g$  and it is beyond the scope of this paper to make an historical review for this statistical problem.

For the sake of simplicity, one shall assume throughout this section that  $\mathbb{X} = \mathbb{R}$ ,  $A = [0, 1]^2$ . The transition density is of the form  $s(x, y) = \varphi(y - g(x))$  where  $\varphi$  is the density of  $\varepsilon_0$ . Since  $g$  and  $\varphi$  are both unknown, this suggests us to consider the class

$$\mathcal{F} = \bigcup_{\sigma > 0} \{f, \exists \phi \in \mathcal{H}^\sigma(\mathbb{R}), \exists g \in \mathcal{B}([0, 1]), \|g\|_\infty < \infty, \forall x, y \in [0, 1], f(x, y) = \phi(y - g(x))\}.$$



A family  $\mathbb{V}$  of linear spaces possessing good approximation properties with respect to the functions of  $\mathcal{F}$  can be built by using Section 6.2 of Baraud and Birgé (2011). Precisely, we prove the following.

**Corollary 4.** *Suppose that Assumption 3 holds with  $\mathbb{X} = \mathbb{R}$ ,  $A = [0, 1]^2$  and with  $\nu \otimes \mu$  the Lebesgue measure on  $\mathbb{R}^2$ . Assume that  $\sqrt{s}|_A$  belongs to  $\mathcal{F}$ . Let  $\sigma > 0$ ,  $p \in (0, +\infty]$ ,  $\beta > (1/p - 1/2)_+$  be any numbers and  $\phi \in \mathcal{H}^\sigma(\mathbb{R})$ ,  $g \in \mathcal{B}^\beta(\mathbb{L}^p([0, 1]))$ ,  $\|g\|_\infty < \infty$  be any functions such that*

$$\sqrt{s(x, y)} = \phi(y - g(x)) \quad \text{for all } x, y \in [0, 1].$$

*There exists two estimators  $\hat{\phi} \geq 0$  and  $\hat{g}$  such that the estimator  $\hat{s}$  defined by*

$$\hat{s}(x, y) = \left( \hat{\phi}(y - \hat{g}(x)) \right)^2 \mathbb{1}_{[0, 1]^2}(x, y) \quad \text{for all } x, y \in \mathbb{R}$$

*satisfies*

$$CE \left[ H^2(s \mathbb{1}_A, \hat{s}) \right] \leq C'_1 \left( \frac{\log^2 n}{n} \right)^{\frac{2\beta(\sigma \wedge 1)}{2\beta(\sigma \wedge 1) + 1}} + C'_2 \left( \frac{\log n}{n} \right)^{\frac{2\sigma}{2\sigma + 1}}$$

*where  $C > 0$  depends only on  $\kappa, p, \sigma, \beta$ , where  $C'_1$  depends only on  $p, \beta, \sigma, |g|_{p, \beta}, \|g\|_\infty, |\phi|_{\infty, \sigma \wedge 1}$  and where  $C'_2$  depends only on  $\sigma, \|g\|_\infty, |\phi|_{\infty, \sigma}$ . Moreover, the construction of the estimators  $\hat{g}$ ,  $\hat{\phi}$  depends only on the data  $X_0, \dots, X_n$ .*

In particular, if  $\phi$  is very smooth (says  $\sigma \geq \beta \vee 1$ ), the rate of convergence corresponds to the rate of convergence for estimating  $g$  only (up to a logarithmic term).

It is interesting to compare the preceding rate to the one we would obtain under the pure smoothness assumption on  $\sqrt{s}|_A$  but ignoring that  $\sqrt{s}|_A$  belongs to  $\mathcal{F}$ . To do so, we need to specify the regularity of  $\sqrt{s}|_A$ , knowing that of  $\phi$  and  $g$ . This is the purpose of the following lemma.

**Lemma 1.** *Let  $\sigma, \beta > 0$ , and let us define*

$$\theta(\beta, \sigma) = \begin{cases} \beta\sigma & \text{if } \beta, \sigma \leq 1 \\ \beta \wedge \sigma & \text{otherwise.} \end{cases}$$

*Let  $\phi \in \mathcal{H}^\sigma(\mathbb{R})$ ,  $g \in \mathcal{H}^\beta([0, 1])$ . The function  $f$  defined by*

$$f(x, y) = \phi(y - g(x)) \quad \text{for all } x, y \in [0, 1],$$

*belongs to  $\mathcal{H}^{(\theta(\beta, \sigma), \sigma)}([0, 1]^2)$ .*

*Moreover, for all  $\sigma, \beta > 0$ , there exist  $\phi \in \mathcal{H}^\sigma(\mathbb{R})$ ,  $g \in \mathcal{H}^\beta([0, 1])$  such that the function  $f$  defined by*

$$f(x, y) = \phi(y - g(x)) \quad \text{for all } x, y \in [0, 1],$$

*belongs to  $\mathcal{H}^{(a, b)}([0, 1]^2)$  if and only if  $a \leq \theta(\beta, \sigma)$  and  $b \leq \sigma$ .*

This result says that if  $\sqrt{s(x, y)} = \phi(y - g(x))$ , with  $\phi \in \mathcal{H}^\sigma(\mathbb{R})$ ,  $g \in \mathcal{H}^\beta([0, 1])$ , then  $\sqrt{s}$  is Hölderian with regularity  $(\theta(\beta, \sigma), \sigma)$  on  $[0, 1]^2$ , and this regularity cannot be improved in general except in some particular situations. Under such a smoothness assumption, the rate of estimation we would get is  $(\log n/n)^{2\sigma\theta(\beta, \sigma)/(2\sigma\theta(\beta, \sigma) + \theta(\beta, \sigma) + \sigma)}$ . This rate is always slower than the rate obtained under the structural assumption.

**4.5. ARCH model.** Throughout this section, we assume that  $X_{n+1} = g_1(X_n) + g_2(X_n)\varepsilon_n$  where  $g_1, g_2$  are unknown functions and where the  $\varepsilon_n$ 's are unobserved identically distributed random variables. The previous model corresponded to  $g_2 = 1$ . The problem of the estimation of the mean and variance functions  $g_1$  and  $g_2$  was considered in several papers and we refer to Section 1.2 of Comte and Rozenholc (2002) for bibliographical references.

For the sake of simplicity, one assumes that  $\mathbb{X} = \mathbb{R}$  and  $A = [0, 1]^2$ . If  $\varphi$  denotes the density of  $\varepsilon_0$ , the transition density  $s$  is of the form

$$(8) \quad s(x, y) = |g_2(x)|^{-1} \varphi[g_2^{-1}(x)(y - g_1(x))] \quad \text{for all } x, y \in \mathbb{R}.$$

We consider thus the class

$$\begin{aligned} \mathcal{F} = \bigcup_{\sigma > 0} \{ & f, \exists \phi \in \mathcal{H}^\sigma(\mathbb{R}), \exists v_1, v_2 \in \mathcal{B}([0, 1]), \|v_1\|_\infty < \infty, \|v_2\|_\infty < \infty, \\ & \forall x, y \in [0, 1], f(x, y) = \sqrt{|v_2(x)|} \phi(v_2(x)(y - v_1(x))) \} \end{aligned}$$

and apply Theorem 5 with a suitable collection  $\mathbb{V}$  to obtain:

**Corollary 5.** *Suppose that Assumption 3 holds with  $\mathbb{X} = \mathbb{R}$ ,  $A = [0, 1]^2$  and with  $\nu \otimes \mu$  the Lebesgue measure on  $\mathbb{R}^2$ . Assume that  $\sqrt{s}|_A$  belongs to  $\mathcal{F}$ . Let  $\sigma > 0$ ,  $\phi \in \mathcal{B}^\sigma(\mathbb{R})$  and for all  $i \in \{1, 2\}$ , let  $p_i \in (0, +\infty]$ ,  $\beta_i > (1/p_i - 1/2)_+$ ,  $v_i \in \mathcal{B}^{\beta_i}(\mathbb{L}^{p_i}([0, 1]))$ , with  $\|v_i\|_\infty < \infty$  such that*

$$\sqrt{s(x, y)} = \sqrt{|v_2(x)|} \phi(v_2(x)(y - v_1(x))) \quad \text{for all } x, y \in [0, 1].$$

*Let  $p_3 \in (0, +\infty]$  and  $\beta_3 > (1/p_3 - 1/2)_+$  be any numbers such that  $v_3 = \sqrt{|v_2|} \in \mathcal{B}^{\beta_3}(\mathbb{L}^{p_3}([0, 1]))$ . There exists an estimator  $\hat{s}$  such that*

$$C\mathbb{E}[H^2(s, \hat{s})] \leq C'_1 \left( \frac{\log^2 n}{n} \right)^{\frac{2\beta(\sigma \wedge 1)}{2\beta(\sigma \wedge 1) + 1}} + C'_2 \left( \frac{\log n}{n} \right)^{\frac{2\sigma}{2\sigma + 1}}$$

where  $\beta = \max(\beta_1, \beta_2, \beta_3)$ . The constant  $C > 0$  depends only on  $\kappa, \sigma, p_1, p_2, p_3, \beta_1, \beta_2, \beta_3$ ,  $C'_1$  depends only on  $\sigma, \|v_1\|_\infty, \|v_2\|_\infty, \|\varphi\|_\infty, |v_1|_{p_1, \beta_1}, |v_2|_{p_2, \beta_2}, |v_3|_{p_3, \beta_3}, |\varphi|_{\infty, \sigma \wedge 1}$  and  $C'_2$  depends only on  $\sigma, \|v_2\|_\infty, |\varphi|_{\infty, \sigma}$ . Moreover, the construction of the estimator  $\hat{s}$  depends only on the data  $X_0, \dots, X_n$ .

If  $s$  is of the form (8) with  $\varphi, g_1, g_2$  smooth, in the sense that  $\phi = \sqrt{\varphi} \in \mathcal{H}^\sigma(\mathbb{R})$ ,  $v_1 = g_1 \in \mathcal{B}^{\beta_1}(\mathbb{L}^{p_1}([0, 1]))$ ,  $\|v_1\|_\infty < \infty$ ,  $v_2 = g_2^{-1} \in \mathcal{B}^{\beta_2}(\mathbb{L}^{p_2}([0, 1]))$ ,  $\|v_2\|_\infty < \infty$  and  $v_3 = |g_2|^{-1/2} \in \mathcal{B}^{\beta_3}(\mathbb{L}^{p_3}([0, 1]))$ , then  $\sqrt{s}|_A$  belongs to  $\mathcal{F}$ . If  $\phi$  is sufficiently smooth ( $\sigma \geq \beta_1 \vee \beta_2 \vee \beta_3 \vee 1$ ), the rate becomes

$$C''\mathbb{E}[H^2(s, \hat{s})] \leq \max \left( \left( \frac{\log^2 n}{n} \right)^{\frac{2\beta_1}{2\beta_1 + 1}}, \left( \frac{\log^2 n}{n} \right)^{\frac{2\beta_2}{2\beta_2 + 1}}, \left( \frac{\log^2 n}{n} \right)^{\frac{2\beta_3}{2\beta_3 + 1}} \right).$$

Up to a logarithmic term, the first term corresponds to the bound we would get if we could estimate  $g_1$  only. The two other terms correspond to the rate of estimation of  $g_2^{-1}$  and  $|g_2|^{-1/2}$  respectively (up to a logarithmic term).

Note that if  $\beta_2 \in (0, 1)$ , one can always choose  $p_3 = 2p_2$  (with  $p_3 = \infty$  if  $p_2 = \infty$ ),  $\beta_3 = \beta_2/2$ , in which case the rate becomes

$$C'' \mathbb{E} [H^2(s, \hat{s})] \leq \max \left( \left( \frac{\log^2 n}{n} \right)^{\frac{2\beta_1}{2\beta_1+1}}, \left( \frac{\log^2 n}{n} \right)^{\frac{\beta_2}{\beta_2+1}} \right).$$

In some situations however,  $\beta_3$  can be taken larger than  $\beta_2$ .

As in the preceding section, we may use the lemma below to compare this rate with the one we would obtain under smoothness assumptions on  $\sqrt{s}|_A$ .

**Lemma 2.** *Let for all  $\sigma, \beta_1, \beta_2 > 0$ ,*

$$\theta(\beta_1, \beta_2, \sigma) = \begin{cases} (2^{-1}(\beta_2 \wedge 1)) \wedge \sigma \beta_1 \wedge \sigma \beta_2 & \text{if } \sigma \leq 1 \text{ and } \beta_1 \wedge \beta_2 \leq 1 \\ (2^{-1}(\beta_2 \wedge 1)) \wedge \sigma \wedge \beta_1 & \text{otherwise.} \end{cases}$$

Let  $\phi \in \mathcal{H}^\sigma(\mathbb{R})$ ,  $v_1 \in \mathcal{H}^{\beta_1}([0, 1])$ ,  $v_2 \in \mathcal{H}^{\beta_2}([0, 1])$ . The function  $f$  defined by

$$f(x, y) = \sqrt{|v_2(x)|} \phi(v_2(x)(y - v_1(x))) \quad \text{for all } x, y \in [0, 1],$$

belongs to  $\mathcal{H}^{(\theta(\beta_1, \beta_2, \sigma), \sigma)}([0, 1]^2)$ .

Moreover, there exist  $\phi \in \mathcal{H}^\sigma(\mathbb{R})$ ,  $v_1 \in \mathcal{H}^{\beta_1}([0, 1])$ ,  $v_2 \in \mathcal{H}^{\beta_2}([0, 1])$  such that the function  $f$  defined by

$$f(x, y) = \sqrt{|v_2(x)|} \phi(v_2(x)(y - v_1(x))) \quad \text{for all } x, y \in [0, 1],$$

belongs to  $\mathcal{H}^{(a, b)}([0, 1]^2)$  if and only if  $a \leq \theta(\beta_1, \beta_2, \sigma)$  and  $b \leq \sigma$ .

This proposition says that if  $\sqrt{s(x, y)} = \sqrt{|v_2(x)|} \phi(v_2(x)(y - v_1(x)))$ , with  $\phi \in \mathcal{H}^\sigma(\mathbb{R})$ ,  $v_1 \in \mathcal{H}^{\beta_1}([0, 1])$ ,  $v_2 \in \mathcal{H}^{\beta_2}([0, 1])$ ,  $\sqrt{s}|_A$  belongs to  $\mathcal{H}^{(\theta(\beta_1, \beta_2, \sigma), \sigma)}([0, 1]^2)$  and the regularity index of this space cannot be increased in general. By Corollary 3, we would get a rate of order  $(\log n/n)^{2\theta(\beta_1, \beta_2, \sigma)\sigma/(2\theta(\beta_1, \beta_2, \sigma)\sigma + \theta(\beta_1, \beta_2, \sigma) + \sigma)}$ , which is slower than the one given by Corollary 5.

## 5. APPENDIX: IMPLEMENTATION OF THE FIRST PROCEDURE.

In this section, we explain how to construct in practice the estimator of the first procedure. This will lead to the proposition below.

**Proposition 6.** *For all  $L > 0$ ,  $\ell \in \mathbb{N}^*$ , the estimator  $\hat{s} = \hat{s}(L, \ell)$  of Section 2.3 can be built in less than  $C(n\ell d + \ell 4^{(\ell+1)d})$  operations where  $C$  is an universal constant.*

We set for all  $K \in \cup_{m \in \mathcal{M}_\ell} m$ ,

$$\hat{s}_K = \frac{\sum_{i=0}^{n-1} \mathbb{1}_K(X_i, X_{i+1})}{\sum_{i=0}^{n-1} \int_{\mathbb{X}} \mathbb{1}_K(X_i, x) d\mu(x)} \mathbb{1}_K,$$

for all  $K' \in \cup_{m \in \mathcal{M}_\ell} m$ ,

$$F_K(K') = \alpha H^2(\hat{s}_K \mathbb{1}_{K'}, \hat{s}_{K'} \mathbb{1}_K) + T(\hat{s}_K \mathbb{1}_{K'}, \hat{s}_{K'} \mathbb{1}_K),$$

and for all  $m' \in \mathcal{M}_\ell$ ,

$$\gamma_K(m') = \left( \sum_{K' \in m'} F_K(K') \right) - \text{pen}(m' \vee K).$$

We shall find for each cube  $K \in \cup_{m \in \mathcal{M}_\ell} m$ , a partition  $m'_K \in \mathcal{M}_\ell$  such that

$$(9) \quad \gamma_K(m'_K) = \sup_{m' \in \mathcal{M}_\ell} \gamma_K(m').$$

We shall compute then

$$(10) \quad \min_{m \in \mathcal{M}_\ell} \gamma(m) = \min_{m \in \mathcal{M}_\ell} \left\{ \left( \sum_{K \in m} \gamma_K(m'_K) \right) + 2\text{pen}(m) \right\}.$$

We shall find  $m'_K$  by using a slight adaptation of the procedure of Blanchard et al. (2004). Computing (10) is similar. The algorithm we propose is based on the one-to-one correspondence between  $\mathcal{M}_\ell$  and the set  $\mathcal{T}_\ell$  of  $4^d$ -ary trees with depth smaller than  $\ell$ .

**Lemma 3.** *There exists  $\psi_\ell$  a one-to-one map between  $\mathcal{M}_\ell$  and  $\mathcal{T}_\ell$  such that for all  $m \in \mathcal{M}_\ell$ ,  $\psi_\ell(m)$  is a tree whose leaves correspond to the elements of the partition  $m$ .*

The construction of this map may for instance be deduced from Section 3.2.4 of Baraud and Birgé (2009).

We need to introduce some notations. For each tree  $T \in \mathcal{T}_\ell$  and bin  $K''$  of  $T$ , we denote by  $T(K'')$  the subtree of  $T$  rooted in  $K''$ . The set of leaves of  $T(K'')$  is denoted by  $\mathcal{L}(T(K''))$ . We set  $R(K'')$  the tree reduced to its root  $K''$  (i.e.,  $\mathcal{L}(R(K'')) = \{K''\}$ ). For all cube  $K \in \cup_{m \in \mathcal{M}_\ell} m$ , we set

$$\mathcal{L}(T(K'')) \vee K = \{K' \cap K, K' \in \mathcal{L}(T(K'')), K' \cap K \neq \emptyset\}$$

and we define the function  $\mathcal{E}$  by

$$\mathcal{E}(T(K'')) = -|\mathcal{L}(T(K'')) \vee K| + \sum_{K' \in \mathcal{L}(T(K''))} F_K(K').$$

The key point is that computing (9) amounts to finding  $T^*$  such that

$$\mathcal{E}(T^*([0, 1]^{2d})) = \sup_{T \in \mathcal{T}_\ell} \mathcal{E}(T([0, 1]^{2d}))$$

since  $m'_K = \psi_\ell^{-1}(T^*)$ .

We now take advantage of the additivity of the function  $\mathcal{E}$ : if  $T(K'')$  is not reduced to its root, and if  $K''_1, \dots, K''_{4^d}$  are the cubes of  $\cup_{m \in \mathcal{M}_\ell} m$  such that  $K'' = \cup_{i=1}^{4^d} K''_i$ , then,

$$(11) \quad \mathcal{E}(T(K'')) = \sum_{i=1}^{4^d} \mathcal{E}(T(K''_i)).$$

For all cube  $K'' \in \cup_{m \in \mathcal{M}_\ell} m$ , let  $T^*(K'')$  be a tree (rooted in  $K''$ ) such that

$$\mathcal{E}(T^*(K'')) = \sup_{T \in \mathcal{T}_\ell, T \ni K''} \mathcal{E}(T(K'')).$$

Remark that if  $K'' \cap K = \emptyset$ , this supremum is equal to 0, in which case  $T^*(K'')$  will always stand for  $R(K'')$ . In general, we deduce from (11) that

$$(12) \quad \mathcal{E}(T^*(K'')) = \max \left( \mathcal{E}(R(K'')), \sum_{i=1}^{4^d} \mathcal{E}(T^*(K_i'')) \right).$$

Calculating (9) can thus be completed in that way: we start with the sets  $K'' \in \cup_{m \in \mathcal{M}_\ell \setminus \mathcal{M}_{\ell-1}} m$  with  $K'' \cap K \neq \emptyset$  for which the optimal local trees are reduced to their roots. By using relation (12) we find the optimal local trees  $T^*(K'')$  when  $K'' \in \cup_{\mathcal{M}_{\ell-1} \setminus \mathcal{M}_\ell} m$ ,  $K'' \cap K \neq \emptyset$ . Proceeding recursively like this yields to the optimal tree  $T^* = T^*([0, 1]^{2d})$ .

## 6. PROOFS

**6.1. Proof of Proposition 1.** Let us introduce the piecewise constant function

$$(13) \quad \bar{s}_m = \sum_{K \in m} \frac{\sum_{i=0}^{n-1} \mathbb{E} [\mathbb{1}_K(X_i, X_{i+1}) \mid X_i]}{\sum_{i=0}^{n-1} \int_{\mathbb{X}} \mathbb{1}_K(X_i, x) d\mu(x)} \mathbb{1}_K.$$

By using the triangular inequality we can decompose the risk of  $\hat{s}_m$  as follows:

$$\mathbb{E} [H^2(s\mathbb{1}_A, \hat{s}_m)] \leq \left(1 + \frac{2 + \log 2}{2}\right) \mathbb{E} [H^2(s\mathbb{1}_A, \bar{s}_m)] + \left(1 + \frac{2}{2 + \log 2}\right) \mathbb{E} [H^2(\bar{s}_m, \hat{s}_m)].$$

The first term can be bounded from above by  $(4 + \log 2) \mathbb{E} [H^2(s\mathbb{1}_A, V_m)]$  thanks to Lemma 2 of Baraud and Birgé (2009). For the second term, we begin to define for  $K \in m$  the random variable

$$B_K = \left( \sqrt{\sum_{i=0}^{n-1} \mathbb{1}_K(X_i, X_{i+1})} - \sqrt{\sum_{i=0}^{n-1} \mathbb{E} [\mathbb{1}_K(X_i, X_{i+1}) \mid X_i]} \right)^2.$$

Since  $2nH^2(\hat{s}_m, \bar{s}_m) = \sum_{K \in m} B_K$ , we shall bound from above the terms  $\mathbb{E}[B_K]$ . For this purpose, we introduce the stopping time

$$T = \inf \left\{ i \geq 0, \mathbb{E} [\mathbb{1}_K(X_i, X_{i+1}) \mid X_i] \geq \frac{1}{2n} \right\} \wedge (n-1)$$

with respect to the filtration  $\mathcal{F}_n = \sigma(X_0, \dots, X_n)$  generated by the random variables  $X_0, \dots, X_n$ . We set  $\varepsilon = 1 + \log 2 + 2 \log n$  and we decompose  $\mathbb{E}[B_K]$  as follows

$$\begin{aligned}
\mathbb{E}[B_K] &\leq (1 + \varepsilon) \mathbb{E} \left[ \left( \sqrt{\sum_{i=0}^{T-1} \mathbb{1}_K(X_i, X_{i+1})} - \sqrt{\sum_{i=0}^{T-1} \mathbb{E}[\mathbb{1}_K(X_i, X_{i+1}) \mid X_i]} \right)^2 \right] \\
&\quad + (1 + \varepsilon^{-1}) \mathbb{E} \left[ \left( \sqrt{\sum_{i=T}^{n-1} \mathbb{1}_K(X_i, X_{i+1})} - \sqrt{\sum_{i=T}^{n-1} \mathbb{E}[\mathbb{1}_K(X_i, X_{i+1}) \mid X_i]} \right)^2 \right] \\
&\leq 2(1 + \varepsilon) \mathbb{E} \left[ \sum_{i=0}^{T-1} \mathbb{E}[\mathbb{1}_K(X_i, X_{i+1}) \mid X_i] \right] \\
(14) \quad &\quad + (1 + \varepsilon^{-1}) \mathbb{E} \left[ \frac{\left( \sum_{i=T}^{n-1} (\mathbb{1}_K(X_i, X_{i+1}) - \mathbb{E}[\mathbb{1}_K(X_i, X_{i+1}) \mid X_i]) \right)^2}{\sum_{i=T}^{n-1} \mathbb{E}[\mathbb{1}_K(X_i, X_{i+1}) \mid X_i]} \right].
\end{aligned}$$

Yet,

$$\mathbb{E} \left[ \sum_{i=0}^{T-1} \mathbb{E}[\mathbb{1}_K(X_i, X_{i+1}) \mid X_i] \right] \leq 1/2,$$

and we control the second term of the right-hand side of inequality (14), by using the claims below.

**Claim 1.** For all  $K \in m$ ,  $j \in \{0, \dots, n\}$ , and  $\mathcal{A}' \in \mathcal{F}_j = \sigma(X_0, \dots, X_j)$ ,

$$\mathbb{E} \left[ \frac{\left( \sum_{i=j}^{n-1} (\mathbb{1}_K(X_i, X_{i+1}) - \mathbb{E}[\mathbb{1}_K(X_i, X_{i+1}) \mid X_i]) \right)^2}{\sum_{i=j}^{n-1} \mathbb{E}[\mathbb{1}_K(X_i, X_{i+1}) \mid X_i]} \mathbb{1}_{\mathcal{A}'} \right] \leq \sum_{k=j}^{n-1} \mathbb{E} \left[ \frac{\mathbb{E}[\mathbb{1}_K(X_k, X_{k+1}) \mid X_k]}{\sum_{i=j}^k \mathbb{E}[\mathbb{1}_K(X_i, X_{i+1}) \mid X_i]} \mathbb{1}_{\mathcal{A}'} \right].$$

*Proof of Claim 1.* Let us define the random variables

$$Y_{n-1}(K) = \sum_{i=j}^{n-1} (\mathbb{1}_K(X_i, X_{i+1}) - \mathbb{E}[\mathbb{1}_K(X_i, X_{i+1}) \mid X_i]) \text{ and } Z_n(K) = \frac{Y_{n-1}^2(K)}{\sum_{i=j}^{n-1} \mathbb{E}[\mathbb{1}_K(X_i, X_{i+1}) \mid X_i]}.$$

We have

$$\begin{aligned}
\mathbb{E}[Z_{n+1}(K) \mid \mathcal{F}_n] &= \frac{\mathbb{E} \left( [Y_{n-1}(K) + (\mathbb{1}_K(X_n, X_{n+1}) - \mathbb{E}[\mathbb{1}_K(X_n, X_{n+1}) \mid X_n])]^2 \mid \mathcal{F}_n \right)}{\sum_{i=j}^n \mathbb{E}[\mathbb{1}_K(X_i, X_{i+1}) \mid X_i]} \\
&= \frac{Y_{n-1}^2(K) + \text{var}(\mathbb{1}_K(X_n, X_{n+1}) \mid X_n)}{\sum_{i=j}^n \mathbb{E}[\mathbb{1}_K(X_i, X_{i+1}) \mid X_i]} \\
&\leq Z_n(K) + \frac{\mathbb{E}[\mathbb{1}_K(X_n, X_{n+1}) \mid X_n]}{\sum_{i=j}^n \mathbb{E}[\mathbb{1}_K(X_i, X_{i+1}) \mid X_i]}.
\end{aligned}$$

Thus, since  $\mathcal{A}'$  is also  $\mathcal{F}_n$ -measurable,

$$\mathbb{E}[Z_{n+1}(K)\mathbb{1}_{\mathcal{A}'}] \leq \mathbb{E}[Z_n(K)\mathbb{1}_{\mathcal{A}'}] + \mathbb{E}\left[\frac{\mathbb{E}[\mathbb{1}_K(X_n, X_{n+1}) \mid X_n]}{\sum_{i=j}^n \mathbb{E}[\mathbb{1}_K(X_i, X_{i+1}) \mid X_i]} \mathbb{1}_{\mathcal{A}'}\right].$$

The result ensues from induction.  $\square$

**Claim 2.** For all sequence  $(u_n)_{n \geq 0}$  in  $[0, 1]$ , and  $j \geq 0$  such that  $u_j \neq 0$ ,

$$\sum_{k=j}^{n-1} \frac{u_k}{\sum_{i=j}^k u_i} \leq 1 + \log n - \log u_j.$$

*Proof of Claim 2.* Let  $f$  be any non-negative continuous function such that  $u_k = \int_k^{k+1} f(t) dt$  whatever  $k \in \mathbb{N}$ . Let  $F$  be the primitive of  $f$  such that  $F(j) = 0$ . Then,

$$\begin{aligned} \sum_{k=j}^{n-1} \frac{u_k}{\sum_{i=j}^k u_i} &\leq 1 + \sum_{k=j+1}^{n-1} \int_k^{k+1} \frac{f(t)}{F(k+1)} dt \\ &\leq 1 + \sum_{k=j+1}^{n-1} \int_k^{k+1} \frac{f(t)}{F(t)} dt \\ &\leq 1 + \log F(n) - \log F(j+1) \\ &\leq 1 + \log \left( \sum_{k=j}^{n-1} u_k \right) - \log u_j. \end{aligned}$$

$\square$

By using Claim 1 with  $\mathcal{A}' = [T = j]$ ,

$$\begin{aligned} \mathbb{E} \left[ \frac{\left( \sum_{i=T}^{n-1} (\mathbb{1}_K(X_i, X_{i+1}) - \mathbb{E}[\mathbb{1}_K(X_i, X_{i+1}) \mid X_i]) \right)^2}{\sum_{i=T}^{n-1} \mathbb{E}[\mathbb{1}_K(X_i, X_{i+1}) \mid X_i]} \right] &\leq \sum_{j=0}^{n-2} \mathbb{E} \left( \sum_{k=j}^{n-1} \frac{\mathbb{E}[\mathbb{1}_K(X_k, X_{k+1}) \mid X_k]}{\sum_{i=j}^k \mathbb{E}[\mathbb{1}_K(X_i, X_{i+1}) \mid X_i]} \mathbb{1}_{T=j} \right) \\ &\quad + \mathbb{E} \left[ \frac{(\mathbb{1}_K(X_{n-1}, X_n) - \mathbb{E}[\mathbb{1}_K(X_{n-1}, X_n) \mid X_{n-1}])^2}{\mathbb{E}[\mathbb{1}_K(X_{n-1}, X_n) \mid X_{n-1}]} \mathbb{1}_{T=n-1} \right]. \end{aligned}$$

Now,

$$\begin{aligned} \mathbb{E} \left[ \frac{(\mathbb{1}_K(X_{n-1}, X_n) - \mathbb{E}[\mathbb{1}_K(X_{n-1}, X_n) \mid X_{n-1}])^2}{\mathbb{E}[\mathbb{1}_K(X_{n-1}, X_n) \mid X_{n-1}]} \mathbb{1}_{T=n-1} \right] &= \mathbb{E} \left[ \frac{\text{var}(\mathbb{1}_K(X_{n-1}, X_n) \mid X_{n-1})}{\mathbb{E}[\mathbb{1}_K(X_{n-1}, X_n) \mid X_{n-1}]} \mathbb{1}_{T=n-1} \right] \\ &\leq \mathbb{P}(T = n-1). \end{aligned}$$

We then use Claim 2 with  $u_k = \mathbb{E}[\mathbb{1}_K(X_k, X_{k+1}) \mid X_k]$  to derive

$$\begin{aligned} \sum_{j=0}^{n-2} \mathbb{E} \left( \sum_{k=j}^{n-1} \frac{\mathbb{E}[\mathbb{1}_K(X_k, X_{k+1}) \mid X_k]}{\sum_{i=j}^k \mathbb{E}[\mathbb{1}_K(X_i, X_{i+1}) \mid X_i]} \mathbb{1}_{T=j} \right) &\leq \sum_{j=0}^{n-2} \mathbb{E}[(1 + \log 2 + 2 \log n) \mathbb{1}_{T=j}] \\ &\leq (1 + \log 2 + 2 \log n) \mathbb{P}(T \neq n-1). \end{aligned}$$

Finally,  $\mathbb{E}[B_K] \leq 4 + 2 \log 2 + 4 \log n$  and hence

$$\mathbb{E}[H^2(\bar{s}_m, \hat{s}_m)] \leq \frac{2 + \log 2 + 2 \log n}{n} |m|,$$

which concludes the proof.  $\square$

**6.2. Proof of Theorem 2.** When  $\ell \leq n$ , the result ensues from the following theorem whose proof is delayed to Section 6.3. In the theorem below, the constant  $L_0 = 90$  can easily be improved but it seems to be difficult to obtain the value  $L_0 = 0.03$  used in practice.

**Theorem 7.** *For all  $L \geq 90$  and  $1 \leq \ell \leq n$ , the estimator  $\hat{s} = \hat{s}(L, \ell)$  satisfies*

$$\forall \xi > 0, \quad \mathbb{P} \left[ CH^2(s\mathbb{1}_A, \hat{s}) \geq \inf_{m \in \mathcal{M}_\ell} (H^2(s\mathbb{1}_A, \hat{s}_m) + \text{pen}(m)) + \xi \right] \leq 3e^{-n\xi}$$

where  $C$  is an universal positive constant.

By integrating the inequality above, there exists  $C' > 0$  such that

$$C' \mathbb{E}[H^2(s\mathbb{1}_A, \hat{s})] \leq \inf_{m \in \mathcal{M}_\ell} \{ \mathbb{E}[H^2(s\mathbb{1}_A, \hat{s}_m)] + \text{pen}(m) \}$$

and the conclusion follows from Proposition 1.

When  $\ell$  is larger than  $n$ , we use the lemma below whose proof is postponed to Section 6.4.

**Lemma 4.** *For all  $L \geq 15$  and  $\ell \geq n + 1$ ,  $\hat{s}(L, \ell) = \hat{s}(L, n)$  and  $\hat{s}(L, \infty) = \hat{s}(L, n)$ .*

Consequently, if  $\ell \geq n + 1$  or  $\ell = \infty$ ,

$$C' \mathbb{E}[H^2(s\mathbb{1}_A, \hat{s})] \leq \inf_{m \in \mathcal{M}_n} \{ \mathbb{E}[H^2(s\mathbb{1}_A, V_m)] + \text{pen}(m) \}.$$

Let  $m^* \in \mathcal{M}_\ell$  such that

$$2 \inf_{m \in \mathcal{M}_\ell} \{ \mathbb{E}[H^2(s\mathbb{1}_A, V_m)] + \text{pen}(m) \} \geq \mathbb{E}[H^2(s\mathbb{1}_A, V_{m^*})] + \text{pen}(m^*).$$

Since

$$\inf_{m \in \mathcal{M}_\ell} \{ \mathbb{E}[H^2(s\mathbb{1}_A, V_m)] + \text{pen}(m) \} \leq \frac{1}{2} + L \frac{\log n}{n},$$

we deduce  $L|m^*| \log(n)/n \leq 1 + 2L \log(n)/n$  and thus  $|m^*| \leq 2 + n/(L \log n) \leq n$ . Remark now that the cardinal of a partition  $m \in \mathcal{M}_\ell \setminus \mathcal{M}_n$  can be lower bounded by

$$|m| \geq 4^d + (4^d - 1)n \geq n + 1.$$

Consequently,  $m^* \in \mathcal{M}_n$  and hence,

$$\inf_{m \in \mathcal{M}_n} \{ \mathbb{E}[H^2(s\mathbb{1}_A, V_m)] + \text{pen}(m) \} \leq 2 \inf_{m \in \mathcal{M}_\ell} \{ \mathbb{E}[H^2(s\mathbb{1}_A, V_m)] + \text{pen}(m) \}$$

which completes the proof.  $\square$



**6.3. Proof of Theorem 7.** The proof of this theorem requires the two following lemmas whose proofs are postponed to Sections 6.3.1 and 6.3.2.

**Lemma 5.** *For all  $m \in \mathcal{M}_\ell$ , there exists a deterministic set  $S_m$  containing  $\hat{s}_m$  such that*

$$\gamma_1(m) = \sup_{m' \in \mathcal{M}_\ell} \{ \alpha H^2(\hat{s}_m, \hat{s}_{m'}) + T(\hat{s}_m, \hat{s}_{m'}) - \text{pen}(m') \} + \text{pen}(m)$$

and

$$\gamma_2(m) = \sup_{\substack{m' \in \mathcal{M}_\ell \\ f' \in S_{m'}}} \{ \alpha H^2(\hat{s}_m, f') + T(\hat{s}_m, f') - \text{pen}(m') \} + 2\text{pen}(m)$$

satisfy

$$\gamma_1(m) \leq \gamma(m) \leq \gamma_2(m).$$

**Lemma 6.** *Set  $\varepsilon = (2 + 3\sqrt{2})/8$ . Under assumptions of Theorem 7, for all  $\xi > 0$ , there exists an event  $\Omega_\xi$  such that  $\mathbb{P}(\Omega_\xi) \geq 1 - 3e^{-n\xi}$  and on which,*

(15) *for all partition  $m \in \mathcal{M}_\ell$ ,*

$$\sup_{\substack{m' \in \mathcal{M}_\ell \\ f' \in S_{m'}}} \{ (1 - \varepsilon) H^2(s\mathbb{1}_A, f') + T(\hat{s}_m, f') - \text{pen}(m') \} \leq (1 + \varepsilon) H^2(s\mathbb{1}_A, \hat{s}_m) + \text{pen}(m) + 22\xi$$

where  $S_{m'}$  is defined in Lemma 5.

*Proof of Theorem 7.* On  $\Omega_\xi$ , for all  $m \in \mathcal{M}_\ell$ ,

$$\sup_{\substack{m' \in \mathcal{M}_\ell \\ f' \in S_{m'}}} \{ (1 - \varepsilon) H^2(s\mathbb{1}_A, f') + T(\hat{s}_m, f') - \text{pen}(m') \} \leq (1 + \varepsilon) H^2(s\mathbb{1}_A, \hat{s}_m) + \text{pen}(m) + 22\xi.$$

If  $T(\hat{s}_m, \hat{s}_{\hat{m}}) + \text{pen}(m) - \text{pen}(\hat{m}) \geq 0$ ,

$$\begin{aligned} \alpha H^2(s\mathbb{1}_A, \hat{s}_{\hat{m}}) &\leq (1 - \varepsilon) H^2(s\mathbb{1}_A, \hat{s}_{\hat{m}}) + T(\hat{s}_m, \hat{s}_{\hat{m}}) - \text{pen}(\hat{m}) + \text{pen}(m) \\ &\leq (1 + \varepsilon) H^2(s\mathbb{1}_A, \hat{s}_m) + 2\text{pen}(m) + 22\xi \end{aligned}$$

since  $\alpha \leq 1 - \varepsilon$  and since  $\hat{s}_{\hat{m}}$  belongs to  $\cup_{m' \in \mathcal{M}_\ell} S_{m'}$ .

If  $T(\hat{s}_m, \hat{s}_{\hat{m}}) + \text{pen}(m) - \text{pen}(\hat{m}) < 0$ ,

$$\begin{aligned} \alpha H^2(\hat{s}_m, \hat{s}_{\hat{m}}) &\leq \alpha H^2(\hat{s}_{\hat{m}}, \hat{s}_m) + T(\hat{s}_{\hat{m}}, \hat{s}_m) - \text{pen}(m) + \text{pen}(\hat{m}) \\ &\leq \sup_{m' \in \mathcal{M}_\ell} \{ \alpha H^2(\hat{s}_{\hat{m}}, \hat{s}_{m'}) + T(\hat{s}_{\hat{m}}, \hat{s}_{m'}) - \text{pen}(m') \} + \text{pen}(\hat{m}) \\ &\leq \gamma_1(\hat{m}). \end{aligned}$$

Consequently, by Lemma 5,

$$\gamma_1(\hat{m}) \leq \gamma(m) + \frac{1}{n} \leq \gamma_2(m) + \frac{1}{n},$$

which implies that

$$\alpha H^2(\hat{s}_m, \hat{s}_{\hat{m}}) \leq \sup_{\substack{m' \in \mathcal{M}_\ell \\ f' \in S_{m'}}} \{ \alpha H^2(\hat{s}_m, f') + T(\hat{s}_m, f') - \text{pen}(m') \} + 2\text{pen}(m) + \frac{1}{n}.$$

With  $v = (1 - \varepsilon)/\alpha - 1 > 0$ ,

$$\begin{aligned}
\alpha H^2(\hat{s}_m, \hat{s}_{\hat{m}}) &\leq (1 + v^{-1}) H^2(\hat{s}_m, s\mathbb{1}_A) \\
&\quad + \sup_{\substack{m' \in \mathcal{M}_\ell \\ f' \in S_{m'}}} \{(1 - \varepsilon) H^2(s\mathbb{1}_A, f') + T(\hat{s}_m, f') - \text{pen}(m')\} + 2\text{pen}(m) + \frac{1}{n} \\
&\leq (1 + v^{-1}) H^2(\hat{s}_m, s\mathbb{1}_A) + [(1 + \varepsilon) H^2(s\mathbb{1}_A, \hat{s}_m) + \text{pen}(m) + 22\xi] + 2\text{pen}(m) + \frac{1}{n} \\
&\leq (2 + \varepsilon + v^{-1}) H^2(\hat{s}_m, s\mathbb{1}_A) + 3\text{pen}(m) + 22\xi + \frac{1}{n}.
\end{aligned}$$

This leads to

$$\begin{aligned}
\alpha H^2(s\mathbb{1}_A, \hat{s}_{\hat{m}}) &\leq 2\alpha H^2(s\mathbb{1}_A, \hat{s}_m) + 2\alpha H^2(\hat{s}_m, \hat{s}_{\hat{m}}) \\
&\leq 2(2 + \alpha + \varepsilon + v^{-1}) H^2(\hat{s}_m, s\mathbb{1}_A) + 6\text{pen}(m) + 44\xi + \frac{2}{n}.
\end{aligned}$$

Finally, we have proved that there exists  $C > 0$ , such that, with probability larger than  $1 - 3e^{-n\xi}$ , for all  $m \in \mathcal{M}_\ell$ ,

$$CH^2(s\mathbb{1}_A, \hat{s}_{\hat{m}}) \leq H^2(\hat{s}_m, s\mathbb{1}_A) + \text{pen}(m) + \xi.$$

This concludes the proof.  $\square$

### 6.3.1. Proof of Lemma 5.

**Claim 3.** *Let, for all  $K \in \cup_{m \in \mathcal{M}_\ell} m$ ,  $K_1, \dots, K_l$  be the cubes of  $\cup_{m \in \mathcal{M}_\ell} m$  such that  $K \subset K_i$  for all  $i \in \{1, \dots, l\}$ . For all  $i \in \{1, \dots, l\}$ , let  $I_i$  and  $J_i$  be the subsets of  $[0, 1]^d$  such that  $K_i = I_i \times J_i$ . Set*

$$S_K = \bigcup_{i=1}^l \left\{ \frac{a}{b\mu(J_i)} \mathbb{1}_K, a \in \{0, \dots, n\}, b \in \{1, \dots, n\} \right\}$$

with the convention  $a/0 = 0$  whatever  $a \in \{0, \dots, n\}$ . Then  $|S_K| \leq \ell n(n+1)$ , and

$$(16) \quad \forall K' \in \cup_{m \in \mathcal{M}_\ell} m, K \subset K', \quad \frac{\sum_{i=0}^{n-1} \mathbb{1}_{K'}(X_i, X_{i+1})}{\sum_{i=0}^{n-1} \int_{\mathbb{X}} \mathbb{1}_{K'}(X_i, x) d\mu(x)} \mathbb{1}_K \in S_K.$$

We then define

$$S_m = \left\{ \sum_{K \in m} f_K, f_K \in S_K \right\}$$

where  $S_K$  is given by the claim above and introduce the random set

$$\hat{S}_m = \left\{ \sum_{K \in m} \hat{s}_{m_K} \mathbb{1}_K, \forall K \in m, m_K \in \mathcal{M}_\ell \right\}.$$

For all  $\hat{f} \in \hat{S}_m$ , we denote by  $m_K(\hat{f})$  any partition of  $\mathcal{M}_\ell$  such that

$$\hat{f}(x) = \hat{s}_{m_K(\hat{f})}(x) \quad \text{for all } x \in K$$

and consider the partition

$$m(\hat{f}) = \bigcup_{K \in m} (m_K(\hat{f}) \vee K).$$

By definition,

$$\gamma(m) = 2\text{pen}(m) + \sum_{K \in m} \sup_{m'_K \in \mathcal{M}_\ell} \left[ \alpha H^2 \left( \hat{s}_m \mathbb{1}_K, \hat{s}_{m'_K} \mathbb{1}_K \right) + T \left( \hat{s}_m \mathbb{1}_K, \hat{s}_{m'_K} \mathbb{1}_K \right) - \text{pen}(m'_K \vee K) \right]$$

and thus

$$\begin{aligned} \gamma(m) &= \sup_{\hat{f} \in \hat{S}_m} \left\{ \sum_{K \in m} \left[ \alpha H^2 \left( \hat{s}_m \mathbb{1}_K, \hat{f} \mathbb{1}_K \right) + T \left( \hat{s}_m \mathbb{1}_K, \hat{f} \mathbb{1}_K \right) - \text{pen}(m_K(\hat{f}) \vee K) \right] \right\} + 2\text{pen}(m) \\ &= \sup_{\hat{f} \in \hat{S}_m} \left\{ \alpha H^2 \left( \hat{s}_m, \hat{f} \right) + T \left( \hat{s}_m, \hat{f} \right) - \text{pen}(m(\hat{f})) \right\} + 2\text{pen}(m). \end{aligned}$$

Now, for all  $m, m' \in \mathcal{M}_\ell$ , the estimator  $\hat{s}_{m'}$ , belongs to  $\hat{S}_m$  with

$$m(\hat{s}_{m'}) = \{K \cap K', (K, K') \in m \times m', K \cap K' \neq \emptyset\},$$

which leads to

$$\gamma(m) \geq \sup_{m' \in \mathcal{M}_\ell} \left\{ \alpha H^2(\hat{s}_m, \hat{s}_{m'}) + T(\hat{s}_m, \hat{s}_{m'}) - \text{pen}(m(\hat{s}_{m'})) \right\} + 2\text{pen}(m).$$

Since  $m(\hat{s}_{m'}) \subset m \cup m'$ ,  $|m(\hat{s}_{m'})| \leq |m| + |m'|$  and  $\gamma_1(m) \leq \gamma(m)$ .

Let us now prove the inequality  $\gamma \leq \gamma_2$ . A function  $\hat{f} \in \hat{S}_m$  is constant on each set of the partition  $m(\hat{f})$ . For  $K' \in m(\hat{f})$ , there exist  $K \in m$ ,  $K'' \in m_K(\hat{f})$  with  $K' \subset K''$  such that

$$\hat{f} \mathbb{1}_{K'} = \frac{\sum_{i=0}^{n-1} \mathbb{1}_{K''}(X_i, X_{i+1})}{\sum_{i=0}^{n-1} \int_{\mathbb{X}} \mathbb{1}_{K''}(X_i, x) d\mu(x)} \mathbb{1}_{K'}.$$

By relation (16),  $\hat{f} \mathbb{1}_{K'} \in S_{K'}$  and thus  $\hat{f} = \sum_{K' \in m(\hat{f})} \hat{f} \mathbb{1}_{K'}$  belongs to  $S_{m(\hat{f})}$ . Consequently,  $\hat{S}_m \subset \cup_{m' \in \mathcal{M}_\ell} S_{m'}$  and the conclusion follows.  $\square$

**6.3.2. Proof of Lemma 6.** We start with the claim below.

**Claim 4.** Let  $\psi$  be the function defined on  $[0, +\infty)^2$  by

$$\psi(x, y) = \frac{1}{\sqrt{2}} \frac{\sqrt{y} - \sqrt{x}}{\sqrt{x+y}} \quad \text{for all } x, y \in [0, +\infty)$$

with the convention  $0/0 = 0$ .

Let, for all  $f, f' \in \mathbb{L}_+^1(\mathbb{X}^2, M)$ , with support included in  $A$ ,  $Z(f, f')$  be the random variable defined by

$$Z(f, f') = \frac{1}{n} \sum_{i=0}^{n-1} \left( \psi(f(X_i, X_{i+1}), f'(X_i, X_{i+1})) - \int_{\mathbb{X}} \psi(f(X_i, y), f'(X_i, y)) (s \mathbb{1}_A)(X_i, y) d\mu(y) \right).$$

Then,

$$(17) \quad \left(1 - \frac{1}{\sqrt{2}}\right) H^2(s \mathbb{1}_A, f') + T(f, f') \leq \left(1 + \frac{1}{\sqrt{2}}\right) H^2(s \mathbb{1}_A, f) + Z(f, f')$$

and

$$(18) \quad \frac{1}{n} \sum_{i=0}^{n-1} \int_{\mathbb{X}} \psi^2(f(X_i, y), f'(X_i, y)) \, d\mu(y) \leq 3 \left( H^2(s\mathbb{1}_A, f) + H^2(s\mathbb{1}_A, f') \right).$$

*Proof.* These inequalities can be obtained by using the same arguments as those used in the proofs of Propositions 2 and 3 of Baraud (2010).  $\square$

We shall prove (15) by applying the following concentration inequality to the random variable  $Z(f, f')$ .

**Claim 5.** *For all  $i \leq n-1$ , let  $\mathcal{F}_i$  be the  $\sigma$ -field generated by the random variables  $X_j$  for  $j \in \{0, \dots, i\}$ . Let  $f_1, \dots, f_n \in \mathbb{L}^1(\mathbb{X}^2, M)$  such that there exists  $b \in \mathbb{R}$  with  $\sup_{x \in \mathbb{X}^2} |f_i(x)| \leq b$  for all  $i \in \{0, \dots, n-1\}$ . Set*

$$S_n = \sum_{i=0}^{n-1} (f_i(X_i, X_{i+1}) - \mathbb{E}[f_i(X_i, X_{i+1}) \mid \mathcal{F}_i])$$

and

$$V_n = \sum_{i=0}^{n-1} \mathbb{E}[f_i^2(X_i, X_{i+1}) \mid \mathcal{F}_i].$$

Then, for all  $\beta > b$  and  $x > 0$

$$\mathbb{P} \left[ S_n \geq \frac{V_n}{2(\beta - b)} + \beta x \right] \leq e^{-x}.$$

*Proof.* By setting  $a^{-1} = 2(\beta - b)$ ,

$$\begin{aligned} \log \mathbb{P} [S_n \geq aV_n + \beta x] &\leq -x + \log \mathbb{E} [\exp(\beta^{-1} S_n - a\beta^{-1} V_n)] \\ &\leq -x + \log \mathbb{E} [\exp(\beta^{-1} S_{n-1} - a\beta^{-1} V_n) \mathbb{E} [\exp(\beta^{-1} (S_n - S_{n-1})) \mid \mathcal{F}_{n-1}]]. \end{aligned}$$

By using Bernstein inequality (Proposition 2.9 of Massart (2003)),

$$\mathbb{E} [\exp(\beta^{-1} (S_n - S_{n-1})) \mid \mathcal{F}_{n-1}] \leq \exp \left( \frac{\beta^{-2} (V_n - V_{n-1})}{2(1 - \beta^{-1} b)} \right)$$

and thus

$$\log \mathbb{P} [S_n \geq aV_n + \beta x] \leq -x + \log \mathbb{E} [\exp(\beta^{-1} S_{n-1} - a\beta^{-1} V_{n-1})].$$

The result follows by induction.  $\square$

*Proof of Lemma 6.* Set  $z = (1 - 1/\sqrt{2})/4$ ,  $\beta = (3/z + \sqrt{2})/2$  and for all  $\xi > 0$ ,

$$\Omega_\xi = \left\{ \sup_{\substack{(f, f') \in S_m \times S_{m'} \\ (m, m') \in \mathcal{M}_\ell^2}} \frac{Z(f, f')}{z(H^2(f, s\mathbb{1}_A) + H^2(f', s\mathbb{1}_A)) + \text{pen}(m) + \text{pen}(m') + \beta\xi} < 1 \right\}.$$

On  $\Omega_\xi$ , for all  $m, m' \in \mathcal{M}_\ell$ ,  $(f, f') \in S_m \times S_{m'}$ ,

$$Z(f, f') \leq z(H^2(f, s\mathbb{1}_A) + H^2(f', s\mathbb{1}_A)) + \text{pen}(m) + \text{pen}(m') + \beta\xi$$

and (15) derives from (17) (with  $\varepsilon = 1/\sqrt{2} + z$ ).

It remains to prove that  $\mathbb{P}(\Omega_\xi^c) \leq 3e^{-n\xi}$ . We have

$$\mathbb{P}(\Omega_\xi^c) \leq \sum_{\substack{(f,f') \in S_m \times S_{m'} \\ (m,m') \in \mathcal{M}_\ell^2}} \mathbb{P}[Z(f, f') \geq z [H^2(s\mathbb{1}_A, f) + H^2(s\mathbb{1}_A, f')] + \text{pen}(m) + \text{pen}(m') + \beta\xi].$$

We apply the concentration inequality given by Claim 5 with  $f_i = \psi(f, f')$ ,  $S_n = nZ(f, f')$  and by using relation (18),

$$V_n = \sum_{i=0}^{n-1} \mathbb{E}[f_i^2(X_i, X_{i+1}) \mid \mathcal{F}_i] \leq 3n(H^2(s\mathbb{1}_A, f) + H^2(s\mathbb{1}_A, f')).$$

We obtain for all  $x > 0$ ,

$$\mathbb{P}\left[Z(f, f') \geq \frac{3}{\sqrt{2}(\beta\sqrt{2} - 1)} [H^2(s\mathbb{1}_A, f) + H^2(s\mathbb{1}_A, f')] + \beta\frac{x}{n}\right] \leq e^{-x}.$$

Note that  $z = 3/(\sqrt{2}(\beta\sqrt{2} - 1))$ . By using the inequality above with

$$\beta\frac{x}{n} = \text{pen}(m) + \text{pen}(m') + \beta\xi$$

we deduce that

$$\mathbb{P}(\Omega_\xi^c) \leq \sum_{\substack{(f,f') \in S_m \times S_{m'} \\ (m,m') \in \mathcal{M}_\ell^2}} e^{-n(\beta^{-1}\text{pen}(m) + \beta^{-1}\text{pen}(m') + \xi)}.$$

Now, by Claim 3, since  $\ell \leq n$ ,  $\log |S_m| \leq 3|m| \log(n+1)$  and thus  $\beta^{-1}\text{pen}(m) \geq (|m| + \log |S_m|)/n$  for all  $m \in \mathcal{M}_\ell$ . Consequently,

$$\begin{aligned} \mathbb{P}(\Omega_\xi^c) &\leq \sum_{\substack{(f,f') \in S_m \times S_{m'} \\ (m,m') \in \mathcal{M}_\ell^2}} e^{-(|m| + \log |S_m| + |m'| + \log |S_{m'}| + n\xi)} \\ &\leq \left( \sum_{m \in \mathcal{M}_\ell} e^{-|m|} \right)^2 e^{-n\xi}. \end{aligned}$$

The conclusion follows from the inequality  $\sum_{m \in \mathcal{M}_\ell} e^{-|m|} \leq \sqrt{3}$  (see Section 3.2.4 of Baraud and Birgé (2009)).  $\square$

**6.4. Proof of Lemma 4.** The lemma follows from the two claims below.

**Claim 6.** Let for each  $m_1, m_2 \in \mathcal{M}_\infty$  and  $K \in m_1$ ,

$$\gamma_K(m_1, m_2) = \alpha H^2(\hat{s}_{m_1} \mathbb{1}_K, \hat{s}_{m_2} \mathbb{1}_K) + T(\hat{s}_{m_1} \mathbb{1}_K, \hat{s}_{m_2} \mathbb{1}_K) - \text{pen}(m_2 \vee K).$$

Then, for all  $\ell \in \mathbb{N}^*$ ,  $\ell \geq n+1$ ,  $m_1 \in \mathcal{M}_\infty$ ,  $K \in m_1$ ,

$$\sup_{m_2 \in \mathcal{M}_\ell} \gamma_K(m_1, m_2) = \sup_{m_2 \in \mathcal{M}_n} \gamma_K(m_1, m_2)$$

and thus

$$\sup_{m_2 \in \mathcal{M}_\infty} \gamma_K(m_1, m_2) = \sup_{m_2 \in \mathcal{M}_n} \gamma_K(m_1, m_2).$$

*Proof.* Let  $m_2^* \in \mathcal{M}_\ell$  such that  $\gamma_K(m_1, m_2^*) = \sup_{m_2 \in \mathcal{M}_\ell} \gamma_K(m_1, m_2)$ . In Section 2, we have defined the collection  $\mathcal{M}_\ell$  of partitions of  $[0, 1]^{2d}$ . Likewise, by using the algorithm of DeVore and Yu (1990), we define the collection  $\mathcal{M}_\ell(K)$  of partitions of  $K$ . Note that  $m_2^* \vee K$  belongs to  $\mathcal{M}_\ell(K)$ . Since  $H^2(\hat{s}_m \mathbb{1}_K, \hat{s}_{m'} \mathbb{1}_K) \leq 1$  and  $|T(\hat{s}_m \mathbb{1}_K, \hat{s}_{m'} \mathbb{1}_K)| \leq 2$ , we have

$$\gamma_K(m_1, m_2^*) \leq 3 - L \frac{|m_2^* \vee K| \log n}{n}.$$

Remark that

$$\gamma_K(m_1, m_2^*) \geq \gamma_K(m_1, \{[0, 1]^{2d}\}) \geq -2 - L \frac{\log n}{n}$$

which leads to

$$|m_2^* \vee K| \leq 1 + \frac{5n}{L \log n} \leq n.$$

This implies that  $m_2^* \vee K$  belongs to  $\mathcal{M}_n(K)$ . There exists  $m_2^\bullet \in \mathcal{M}_n$  such that  $m_2^\bullet \vee K = m_2^* \vee K$  and hence  $\gamma_K(m_1, m_2^\bullet) = \gamma_K(m_1, m_2^*)$  which concludes the proof.  $\square$

**Claim 7.** Set for all  $m \in \mathcal{M}_\infty$  and  $K \in m$ ,

$$\gamma_K(m) = \sup_{m_2 \in \mathcal{M}_\ell} \gamma_K(m, m_2).$$

Then,  $\gamma(m) = 2\text{pen}(m) + \sum_{K \in m} \gamma_K(m)$  and for all  $\ell \in \mathbb{N}^*$ ,  $\ell \geq n + 1$ ,

$$\inf_{m \in \mathcal{M}_\ell} \gamma(m) = \inf_{m \in \mathcal{M}_n} \gamma(m)$$

and thus

$$\inf_{m \in \mathcal{M}_\infty} \gamma(m) = \inf_{m \in \mathcal{M}_n} \gamma(m).$$

*Proof.* Let  $m^* \in \mathcal{M}_\ell$  such that  $\inf_{m \in \mathcal{M}_\ell} \gamma(m) = \gamma(m^*)$ . By Lemma 5,

$$\begin{aligned} \gamma(m^*) &\geq \sup_{m' \in \mathcal{M}_\ell} \left\{ \alpha H^2(\hat{s}_m, \hat{s}_{m'}) + T(\hat{s}_m, \hat{s}_{m'}) - \text{pen}(m') \right\} + L \frac{|m^*| \log n}{n} \\ &\geq \left( -2 - L \frac{\log n}{n} \right) + L \frac{|m^*| \log n}{n} \\ &\geq -2 + L \frac{(|m^*| - 1) \log n}{n}. \end{aligned}$$

Now,

$$\gamma(m^*) \leq \gamma(\{[0, 1]^{2d}\}) \leq 2L \frac{\log n}{n} + 3$$

which implies that

$$|m^*| \leq 3 + \frac{5n}{L \log n} \leq n$$

and thus  $m^* \in \mathcal{M}_n$ .  $\square$

**6.5. Proof of Theorem 3.** Consider the regular partition  $m_{ref}$  of  $[0, 1]^{2d}$  into cubes with side length  $2^{-\ell}$ , that is

$$m_{ref} = \left\{ K_{\ell, \mathbf{l}}, \mathbf{l} = (k, \dots, k), k \in \{1, \dots, 2^\ell\} \right\}$$

where  $K_{\ell, \mathbf{l}}$  is defined in Section 2.2. For all partition  $m \in \mathcal{M}_\ell$ ,  $V_m \subset V_{m_{ref}}$ . Set

$$\Omega_{eq} = [\forall g_1, g_2 \in V_{m_{ref}}, h^2(g_1, g_2) \leq 11H^2(g_1, g_2)]$$

and define  $\bar{s}_m$  an element of  $V_m$  such that  $h^2(s\mathbb{1}_A, \bar{s}_m) = h^2(s\mathbb{1}_A, V_m)$ .

For all  $m \in \mathcal{M}_\ell$ ,

$$\begin{aligned} \mathbb{E} [h^2(s\mathbb{1}_A, \hat{s}_{\hat{m}})] &\leq \mathbb{E} [h^2(s\mathbb{1}_A, \hat{s}_{\hat{m}}) \mathbb{1}_{\Omega_{eq}}] + \mathbb{E} [h^2(s\mathbb{1}_A, \hat{s}_{\hat{m}}) \mathbb{1}_{\Omega_{eq}^c}] \\ &\leq 2\mathbb{E} [h^2(s\mathbb{1}_A, \bar{s}_m) \mathbb{1}_{\Omega_{eq}}] + 2\mathbb{E} [h^2(\bar{s}_m, \hat{s}_{\hat{m}}) \mathbb{1}_{\Omega_{eq}}] + \mathbb{E} [h^2(s\mathbb{1}_A, \hat{s}_{\hat{m}}) \mathbb{1}_{\Omega_{eq}^c}] \\ &\leq 2\mathbb{E} [h^2(s\mathbb{1}_A, \bar{s}_m) \mathbb{1}_{\Omega_{eq}}] + 22\mathbb{E} [H^2(\bar{s}_m, \hat{s}_{\hat{m}}) \mathbb{1}_{\Omega_{eq}}] + \mathbb{E} [h^2(s\mathbb{1}_A, \hat{s}_{\hat{m}}) \mathbb{1}_{\Omega_{eq}^c}] \\ &\leq 2\mathbb{E} [h^2(s\mathbb{1}_A, \bar{s}_m) \mathbb{1}_{\Omega_{eq}}] + 44\mathbb{E} [H^2(s\mathbb{1}_A, \bar{s}_m) \mathbb{1}_{\Omega_{eq}}] + 44\mathbb{E} [H^2(\hat{s}_{\hat{m}}, s\mathbb{1}_A) \mathbb{1}_{\Omega_{eq}}] \\ &\quad + \mathbb{E} [h^2(s\mathbb{1}_A, \hat{s}_{\hat{m}}) \mathbb{1}_{\Omega_{eq}^c}]. \end{aligned}$$

Now,  $h^2(s\mathbb{1}_A, \bar{s}_m) = \mathbb{E}[H^2(s\mathbb{1}_A, \bar{s}_m)] = h^2(s\mathbb{1}_A, V_m)$  and

$$\begin{aligned} \mathbb{E} [h^2(s\mathbb{1}_A, \hat{s}_{\hat{m}}) \mathbb{1}_{\Omega_{eq}^c}] &\leq 2\mathbb{E} [(h^2(s, 0) + h^2(\hat{s}_{\hat{m}}, 0)) \mathbb{1}_{\Omega_{eq}^c}] \\ &\leq \mathbb{E} \left[ \left( 1 + 2 \sup_{m \in \mathcal{M}_\ell} h^2(\hat{s}_m, 0) \right) \mathbb{1}_{\Omega_{eq}^c} \right]. \end{aligned}$$

Let for all  $K \in m$ ,  $I_K$  and  $J_K$  be the subsets of  $[0, 1]^d$  such that  $K = I_K \times J_K$ . Then,

$$2h^2(\hat{s}_m, 0) = \sum_{K \in m} \frac{\sum_{i=0}^{n-1} \mathbb{1}_{I_K}(X_i) \mathbb{1}_{J_K}(X_{i+1})}{\sum_{i=0}^{n-1} \mathbb{1}_{I_K}(X_i)} \int_{I_K} \varphi(x) dx \leq |m|.$$

Since  $m \subset m_{ref}$ ,  $|m| \leq |m_{ref}| = 4^{\ell d}$  and thus,

$$C' \mathbb{E} [h^2(s\mathbb{1}_A, \hat{s}_{\hat{m}})] \leq \inf_{m \in \mathcal{M}_\ell} \{h^2(s\mathbb{1}_A, V_m) + \text{pen}(m)\} + 4^{\ell d} \mathbb{P}(\Omega_{eq}^c)$$

for some universal constant  $C' > 0$ .

We now bound from above the term  $\mathbb{P}(\Omega_{eq}^c)$ . We denote by  $\mathbf{I}_{ref}$  the regular partition of  $[0, 1]^d$  into cubes with side length  $2^{-\ell}$ . Remark that

$$\begin{aligned} \mathbb{P}(\Omega_{eq}^c) &\leq \mathbb{P} \left[ \exists I \in \mathbf{I}_{ref}, \mathbb{P}(X_1 \in I) \geq \frac{11}{n} \sum_{i=0}^{n-1} \mathbb{1}_I(X_i) \right] \\ &\leq 2^{\ell d} \sup_{I \in \mathbf{I}_{ref}} \mathbb{P} \left[ \frac{1}{n} \sum_{i=0}^{n-1} (\mathbb{1}_I(X_i) - \mathbb{P}(X_i \in I)) \leq -\frac{10}{11} \mathbb{P}(X_1 \in I) \right]. \end{aligned}$$

We use the following Bennett-type inequality for  $\beta$ -mixing random variables (with  $f = -\mathbb{1}_I$ ,  $v = \mathbb{P}(X_1 \in I)$ ,  $c = 0$ ,  $\xi = 10/11\mathbb{P}(X_1 \in I)$ ).

**Proposition 8.** *Let  $(X_i)_{i \geq 1}$  be a stationary sequence of random vectors with values in  $\mathbb{R}^d$ , and let  $f$  be a real-valued function on  $\mathbb{R}^d$  bounded from above by  $c \geq 0$  such that  $v = \mathbb{E}[f(X_i)^2] < \infty$ .*

*Then, for all  $q \in \{1, \dots, n\}$  and  $\xi > 0$ ,*

$$\mathbb{P} \left( \frac{1}{n} \sum_{i=1}^n (f(X_i) - \mathbb{E}(f(X_i))) > \xi \right) \leq 2 \exp \left( -\frac{n\xi^2}{8q(v + c\xi/6)} \right) + 3n\beta_q/q.$$

We then have for all  $I \in \mathbf{I}_{ref}$ ,

$$\begin{aligned} \mathbb{P} \left[ \frac{1}{n} \sum_{i=0}^{n-1} (\mathbb{1}_I(X_i) - \mathbb{P}(X_i \in I)) \leq -\frac{10}{11} \mathbb{P}(X_1 \in I) \right] &\leq 3 \inf_{1 \leq q \leq n} \left\{ \exp \left( -\frac{25n\mathbb{P}(X_1 \in I)}{242q} \right) + n\beta_q/q \right\} \\ &\leq 3 \inf_{1 \leq q \leq n} \left\{ \exp \left( -\frac{n\kappa_0}{10q2^{\ell d}} \right) + n\beta_q/q \right\} \end{aligned}$$

which concludes the proof.  $\square$

*Proof of Proposition 8.* Let  $l$  be the smallest integer larger than  $n/(2q)$ . We derive from Berbee's lemma and more precisely from Viennet (1997) (page 484) that there exist  $X_1^*, \dots, X_{2lq}^*$  such that

- For  $j = 1, \dots, l$ , the random vectors

$$\mathbf{X}_{j,1} = (X_{2(j-1)q+1}, \dots, X_{2(j-1)q+q}) \quad \text{and} \quad \mathbf{X}_{j,1}^* = (X_{2(j-1)q+1}^*, \dots, X_{2(j-1)q+q}^*)$$

have the same distribution, and so have the random vectors

$$\mathbf{X}_{j,2} = (X_{2(j-1)q+q+1}, \dots, X_{2jq}) \quad \text{and} \quad \mathbf{X}_{j,2}^* = (X_{2(j-1)q+q+1}^*, \dots, X_{2jq}^*).$$

- The random vectors  $\mathbf{X}_{1,1}^*, \dots, \mathbf{X}_{l,1}^*$  are independent. The random vectors  $\mathbf{X}_{1,2}^*, \dots, \mathbf{X}_{l,2}^*$  are also independent.
- The event

$$\Omega^* = \bigcap_{1 \leq j \leq l} ([\mathbf{X}_{j,1} \neq \mathbf{X}_{j,1}^*] \cap [\mathbf{X}_{j,2} \neq \mathbf{X}_{j,2}^*])$$

satisfies  $\mathbb{P}[(\Omega^*)^c] \leq 2l\beta_q$ .

We set  $g_i(x) = f(x)$  if  $i \leq n$  and  $g_i(x) = 0$  otherwise. For  $j \in \{1, \dots, l\}$ , we set

$$g'_{j,1}(x_1, \dots, x_q) = \sum_{i=1}^q g_{2(j-1)q+i}(x_i) \quad \text{and} \quad g'_{j,2}(x_1, \dots, x_q) = \sum_{i=1}^q g_{2(j-1)q+q+i}(x_i).$$

Then,

$$\begin{aligned} \mathbb{P} \left[ \left( \frac{1}{n} \sum_{i=1}^n (g_i(X_i) - \mathbb{E}[g_i(X_i)]) > \xi \right) \cap \Omega^* \right] &\leq \mathbb{P} \left( \sum_{j=1}^l (g'_{j,1}(\mathbf{X}_{j,1}^*) - \mathbb{E}[g'_{j,1}(\mathbf{X}_{j,1}^*)]) > n\xi/2 \right) \\ &\quad + \mathbb{P} \left( \sum_{j=1}^l (g'_{j,2}(\mathbf{X}_{j,2}^*) - \mathbb{E}[g'_{j,2}(\mathbf{X}_{j,2}^*)]) > n\xi/2 \right) \\ &\leq 2 \exp \left( -\frac{n^2\xi^2}{8q(nv + cn\xi/6)} \right) \end{aligned}$$



by using Proposition 2.8 of Massart (2003).  $\square$

**6.6. Proof of Corollary 2.** The corollary ensues from the claim below and Theorem 2 of Baraud and Birgé (2009).

**Claim 8.** *Under Assumption 2, for all  $\ell \in \mathbb{N}^*$  such that  $2^{\ell d} \geq n$ ,*

$$\inf_{m \in \mathcal{M}_\ell} \left\{ d_2^2(\sqrt{s}|_A, V_m) + \frac{|m| \log n}{n} \right\} \leq 4 \inf_{m \in \mathcal{M}_\infty} \left\{ d_2^2(\sqrt{s}|_A, V_m) + \frac{|m| \log n}{n} \right\}.$$

*Proof.* For all partition  $m \in \mathcal{M}_\infty$  and cube  $K \in m$ , we denote by  $I_K$  and  $J_K$  the cubes of  $[0, 1]^d$  such that  $K = I_K \times J_K$  and set

$$\bar{s}_m = \sum_{K \in m} \frac{\int_K s(x, y) dx dy}{\mu \otimes \mu(K)} \mathbb{1}_K.$$

In this paper,  $d_2$  stands for the standard euclidean distance of  $\mathbb{L}^2([0, 1]^{2d}, \mu \otimes \mu)$ . In this proof, we make a small abuse of notations by denoting by  $d_2$  the standard euclidean distance of  $\mathbb{L}^2(\mathbb{R}^{2d}, \mu \otimes \mu)$ .

Let  $m^*$  be a partition of  $\mathcal{M}_\infty$  such that

$$2 \inf_{m \in \mathcal{M}_\infty} \left\{ d_2^2(\sqrt{s} \mathbb{1}_A, V_m) + \frac{|m| \log n}{n} \right\} \geq d_2^2(\sqrt{s} \mathbb{1}_A, V_{m^*}) + \frac{|m^*| \log n}{n}.$$

Let  $\mathcal{C}$  be the collection  $\mathcal{C} = \{K \in m^*, \mu(I_K) \geq 2^{-\ell d}\}$  and let  $m^\bullet$  be a partition of  $\mathcal{M}_\ell$  containing  $\mathcal{C}$  such that

$$|m^\bullet| = \inf\{|m|, m \in \mathcal{M}_\ell \text{ such that } m \ni \mathcal{C}\}.$$

Let  $A^\bullet$  be the set defined by  $A^\bullet = \cup_{K \in m^\bullet} K$  and  $V_{m^\bullet} = \{f \mathbb{1}_{A^\bullet}, f \in V_{m^*}\}$ .

We have,

$$d_2^2(\sqrt{s} \mathbb{1}_A, V_{m^\bullet}) \leq d_2^2(\sqrt{s} \mathbb{1}_{A^\bullet}, V_{m^\bullet}^\bullet) + d_2^2(\sqrt{s} \mathbb{1}_{A \cap (A^\bullet)^c}, 0)$$

and

$$d_2^2(\sqrt{s} \mathbb{1}_{A^\bullet}, V_{m^\bullet}^\bullet) \leq d_2^2(\sqrt{s} \mathbb{1}_{A^\bullet}, \sqrt{\bar{s}_{m^\bullet}} \mathbb{1}_{A^\bullet}) \leq d_2^2(\sqrt{s} \mathbb{1}_A, \sqrt{\bar{s}_{m^*}}).$$

By using Lemma 2 of Baraud and Birgé (2009),  $d_2^2(\sqrt{s} \mathbb{1}_A, \sqrt{\bar{s}_{m^*}}) \leq 2d_2^2(\sqrt{s} \mathbb{1}_A, V_{m^*})$  which shows that

$$d_2^2(\sqrt{s} \mathbb{1}_A, V_{m^\bullet}) \leq 2d_2^2(\sqrt{s} \mathbb{1}_A, V_{m^*}) + d_2^2(\sqrt{s} \mathbb{1}_{A \cap (A^\bullet)^c}, 0).$$

Now,

$$\begin{aligned} d_2^2(\sqrt{s} \mathbb{1}_{A \cap (A^\bullet)^c}, 0) &\leq \sum_{K \in m^* \setminus \mathcal{C}} \int_{I_K} \left( \int_{\mathbb{R}^d} s(x, y) dy \right) dx \\ &\leq \sum_{K \in m^* \setminus \mathcal{C}} \mu(I_K) \leq 2^{-\ell d} |m^*|. \end{aligned}$$

Since  $|m^\bullet| \leq |m^*|$ , we have

$$d_2^2(\sqrt{s} \mathbb{1}_A, V_{m^\bullet}) + \frac{|m^\bullet| \log n}{n} \leq 2d_2^2(\sqrt{s} \mathbb{1}_A, V_{m^*}) + \frac{(1 + \log n)|m^*|}{n}$$

which proves the claim.  $\square$

**6.7. Rates of convergences for  $h$ .** We prove the result only for geometrically  $\beta$ -mixing chains (the proof for arithmetically  $\beta$ -mixing chains being similar). We use the claim below whose proof is the same than the one of Claim 8.

**Claim 9.** *Under Assumption 2, for all  $\ell \in \mathbb{N}^*$  such that  $2^{\ell d} \geq n/\log^3 n$ ,*

$$\inf_{m \in \mathcal{M}_\ell} \left\{ h^2(s\mathbb{1}_A, V_m) + \frac{|m| \log n}{n} \right\} \leq 4 \inf_{m \in \mathcal{M}_\infty} \left\{ h^2(s\mathbb{1}_A, V_m) + \frac{|m| \log^3 n}{n} \right\}.$$

By using this claim and Theorem 1 of Akakpo (2012),

$$(19) \quad C\mathbb{E} \left[ h^2(s\mathbb{1}_A, \hat{s}_{\hat{m}}) \right] \leq |\sqrt{s}|_A \Big|_{p,\sigma}^{\frac{2d}{d+\sigma}} \left( \frac{\log^3 n}{n} \right)^{\frac{\sigma}{\sigma+d}} + \frac{\log^3 n}{n} + \frac{R_n(\ell)}{n}$$

and by using Theorem 2 of Akakpo (2012),

$$C\mathbb{E} \left[ h^2(s\mathbb{1}_A, \hat{s}_{\hat{m}}) \right] \leq |\sqrt{s}|_A \Big|_{p,\sigma}^{\frac{2d}{d+\sigma}} \left( \frac{\log n}{n} + 2^{-2\ell d\theta} \right)^{\frac{\sigma}{\sigma+d}} + \frac{\log n}{n} + \frac{R_n(\ell)}{n}$$

where  $C > 0$  depends only on  $\kappa, \sigma, d, p$  and where

$$\theta = \frac{d+\sigma}{\sigma} \left( \frac{\sigma}{d} - 2 \left( \frac{1}{p} - \frac{1}{2} \right)_+ \right).$$

If  $\sigma > \sigma_1(p, d)$  then  $\theta > 1/2$ . There exists thus  $n_0$  (depending only on  $\theta$ ), such that if  $n \geq n_0$ ,  $2^{-2\ell d\theta} \leq \log n/n$ , and hence

$$C'\mathbb{E} \left[ h^2(s\mathbb{1}_A, \hat{s}_{\hat{m}}) \right] \leq |\sqrt{s}|_A \Big|_{p,\sigma}^{\frac{2d}{d+\sigma}} \left( \frac{\log n}{n} \right)^{\frac{\sigma}{\sigma+d}} + \frac{\log n}{n} + \frac{R_n(\ell)}{n}.$$

If  $n \leq n_0$ , we deduce from (19),

$$\begin{aligned} C\mathbb{E} \left[ h^2(s\mathbb{1}_A, \hat{s}_{\hat{m}}) \right] &\leq |\sqrt{s}|_A \Big|_{p,\sigma}^{\frac{2d}{d+\sigma}} \left( \frac{\log^3 n_0}{n} \right)^{\frac{\sigma}{\sigma+d}} + \frac{\log^3 n_0}{n} + \frac{R_n(\ell)}{n} \\ &\leq C'' \left[ |\sqrt{s}|_A \Big|_{p,\sigma}^{\frac{2d}{d+\sigma}} \left( \frac{\log n}{n} \right)^{\frac{\sigma}{\sigma+d}} + \frac{\log n}{n} + \frac{R_n(\ell)}{n} \right] \end{aligned}$$

where  $C''$  depends only on  $\sigma, d, p$ . The conclusion ensues from the fact that  $R_n(\ell)$  is upper-bounded by a constant depending only on  $\kappa_0, b_1$ .  $\square$

**6.8. Proof of Proposition 4.** We shall use the following lemma whose proof is similar to the one of Lemma 6.

**Lemma 7.** *Set  $\varepsilon = (2 + 3\sqrt{2})/8$ . Under assumptions of Proposition 4, there exists an universal constant  $L_0 > 0$  such that for all  $L \geq L_0$  and  $\xi > 0$ ,*

$$\forall f, f' \in S, \quad (1 - \varepsilon) H^2(s\mathbb{1}_A, f') + T(f, f') \leq (1 + \varepsilon) H^2(s\mathbb{1}_A, f) + L \frac{\Delta_S(f) + \Delta_S(f')}{n} + 22\xi$$

with probability larger than  $1 - e^{-n\xi}$ .

*Proof of Proposition 4.* By using the lemma above, with probability larger than  $1 - e^{-n\xi}$ , for all  $f \in S$ ,

$$\sup_{f' \in S} \left\{ (1 - \varepsilon) H^2(s\mathbb{1}_A, f') + T(f, f') - L \frac{\Delta_S(f')}{n} \right\} \leq (1 + \varepsilon) H^2(s\mathbb{1}_A, f) + L \frac{\Delta_S(f)}{n} + 22\xi.$$

Thus, if  $T(f, \hat{f}) + L \frac{\Delta_S(f)}{n} - L \frac{\Delta_S(\hat{f})}{n} \geq 0$ ,

$$\begin{aligned} \alpha H^2(s\mathbb{1}_A, \hat{f}) &\leq (1 - \varepsilon) H^2(s\mathbb{1}_A, \hat{f}) + T(f, \hat{f}) - L \frac{\Delta_S(\hat{f})}{n} + L \frac{\Delta_S(f)}{n} \\ &\leq (1 + \varepsilon) H^2(s\mathbb{1}_A, f) + 2L \frac{\Delta_S(f)}{n} + 22\xi. \end{aligned}$$

If  $T(f, \hat{f}) + L \frac{\Delta_S(f)}{n} - L \frac{\Delta_S(\hat{f})}{n} < 0$ ,

$$\begin{aligned} \alpha H^2(f, \hat{f}) &\leq \alpha H^2(\hat{f}, f) + T(\hat{f}, f) - L \frac{\Delta_S(f)}{n} + L \frac{\Delta_S(\hat{f})}{n} \\ &\leq \sup_{f' \in S} \left\{ \alpha H^2(\hat{f}, f') + T(\hat{f}, f') - L \frac{\Delta_S(f')}{n} \right\} + L \frac{\Delta_S(\hat{f})}{n} \\ &\leq \wp(\hat{f}) \\ &\leq \wp(f) + \frac{1}{n} \\ &\leq \sup_{f' \in S} \left\{ \alpha H^2(f, f') + T(f, f') - L \frac{\Delta_S(f')}{n} \right\} + L \frac{\Delta_S(f)}{n} + \frac{1}{n}. \end{aligned}$$

With  $v = (1 - \varepsilon)/\alpha - 1 > 0$ ,

$$\begin{aligned} \alpha H^2(f, \hat{f}) &\leq (1 + v^{-1}) H^2(f, s\mathbb{1}_A) \\ &\quad + \sup_{f' \in S} \left\{ (1 - \varepsilon) H^2(s\mathbb{1}_A, f') + T(f, f') - L \frac{\Delta_S(f')}{n} \right\} + L \frac{\Delta_S(f)}{n} + \frac{1}{n} \\ &\leq (1 + v^{-1}) H^2(f, s\mathbb{1}_A) + \left[ (1 + \varepsilon) H^2(s\mathbb{1}_A, f) + L \frac{\Delta_S(f)}{n} + 22\xi \right] + L \frac{\Delta_S(f)}{n} + \frac{1}{n} \\ &\leq (2 + \varepsilon + v^{-1}) H^2(f, s\mathbb{1}_A) + 2L \frac{\Delta_S(f)}{n} + 22\xi + \frac{1}{n}. \end{aligned}$$

This leads to,

$$\begin{aligned} \alpha H^2(s\mathbb{1}_A, \hat{f}) &\leq 2\alpha H^2(s\mathbb{1}_A, f) + 2\alpha H^2(f, \hat{f}) \\ &\leq 2(2 + \alpha + \varepsilon + v^{-1}) H^2(f, s\mathbb{1}_A) + 4L \frac{\Delta_S(f)}{n} + 44\xi + \frac{2}{n}. \end{aligned}$$

Finally, we have proved that there exists  $C > 0$ , such that, with probability larger than  $1 - e^{-n\xi}$ , for all  $f \in S$ ,

$$CH^2(s\mathbb{1}_A, \hat{f}) \leq H^2(f, s\mathbb{1}_A) + L \frac{\Delta_S(f)}{n} + \xi.$$

The conclusion follows.  $\square$

**6.9. Proof of Corollary 4.** Throughout this proof, the distance associated to the supremum norm  $\|\cdot\|_\infty$  is denoted by  $d_\infty$ . We shall use the following lemma (the first part may be deduced from the work of Akakpo (2012) whereas the second part may be deduced from results in Dahmen et al. (1980)).

**Lemma 8.** *There exists a collection  $\mathbb{W}$  of (finite dimensional) linear spaces such that for all  $p \in (0, +\infty]$ ,  $\beta > (1/p - 1/2)_+$  and  $f \in \mathcal{B}^\beta(\mathbb{L}^p([0, 1]))$ ,  $L > 0$ ,  $\tau > 0$ ,  $\sigma > 0$ ,*

$$C \inf_{W \in \mathbb{W}} \{L^2 d_2^{2\sigma}(g, W) + (\dim W)\tau\} \leq (L|g|_{p,\beta}^\sigma)^{\frac{2}{2\sigma\beta+1}} \tau^{\frac{2\sigma\beta}{2\sigma\beta+1}} + \tau$$

where  $C > 0$  depends only on  $p, \beta$ . Moreover, for all  $\beta > 0$ ,  $f \in \mathcal{H}^\beta([0, 1])$ ,  $L > 0$ ,  $\tau > 0$ ,  $\sigma > 0$ ,

$$C' \inf_{W \in \mathbb{W}} \{L^2 d_\infty^{2\sigma}(g, W) + (\dim W)\tau\} \leq (L|g|_{\infty,\beta}^\sigma)^{\frac{2}{2\sigma\beta+1}} \tau^{\frac{2\sigma\beta}{2\sigma\beta+1}} + \tau$$

where  $C' > 0$  depends only on  $\beta$ .

Let us define

$$u(x, y) = \frac{y - g(x)}{1 + \|g\|_\infty} \quad \text{and} \quad \Phi(x) = \phi((1 + \|g\|_\infty)x) \quad \text{for all } x, y \in [0, 1].$$

Let  $\mathbb{W}$  be the family of linear spaces given by the lemma above. Define, for all  $W \in \mathbb{W}$ , the linear space

$$T_W = \{(x, y) \mapsto a(y - f(x)), a \in \mathbb{R}, f \in W\}$$

and  $\mathbb{T} = \{T_W, W \in \mathbb{W}\}$ . Since  $\Phi$  belongs to  $\mathcal{H}^\sigma([0, 1])$ , we deduce from Corollary 1 of Baraud and Birgé (2011) and from our Theorem 5 that there exists an estimator  $\hat{s}$  such that

$$C\mathbb{E}[H^2(s, \hat{s})] \leq \inf_{T \in \mathbb{T}} \left\{ |\Phi|_{\infty, \sigma \wedge 1}^2 d_2^{2(\sigma \wedge 1)}(u, T) + (\dim T)\tau_n \right\} + \inf_{W \in \mathbb{W}} \left\{ d_\infty^2(\Phi, W) + (\dim W) \frac{\log n}{n} \right\}$$

where  $C > 0$  depends on  $\sigma, \kappa$  and where

$$\tau_n = (\log n \vee \log(|\Phi|_{\infty, \sigma \wedge 1})) \frac{\log n}{n}.$$

Now,

$$\inf_{T \in \mathbb{T}} \left\{ |\Phi|_{\infty, \sigma \wedge 1}^2 d_2^{2(\sigma \wedge 1)}(u, T) + (\dim T)\tau_n \right\} \leq \inf_{W \in \mathbb{W}} \left\{ |\phi|_{\infty, \sigma \wedge 1}^2 d_2^{2(\sigma \wedge 1)}(g, W) + (\dim W + 1)\tau_n \right\}$$

and the conclusion follows from the lemma above.  $\square$

**6.10. Proof of Lemma 1.** The first part of the lemma may be deduced from Proposition 4 of Baraud and Birgé (2011). For the second part, we shall build  $\phi' \in \mathcal{H}^\sigma(\mathbb{R})$  such that  $\phi'|_{[0,1]} \notin \cup_{b>\sigma} \mathcal{H}^b([0,1])$  and  $g' \in \mathcal{H}^\beta([0,1])$  such that  $g'(0) = 0$  and

$$\phi' \circ g' \in \mathcal{H}^{\theta(\beta, \sigma)}([0, 1]) \setminus \cup_{b>\theta(\beta, \sigma)} \mathcal{H}^b([0, 1]).$$

By setting  $\phi = \phi'$  and  $g = -g'$ , the function  $f$  defined by

$$f(x, y) = \phi'(y - (-g'(x))) \quad \text{for all } x, y \in [0, 1],$$

is suitable since  $f(x, 0) = \phi' \circ g'(x)$  and  $f(0, y) = \phi'(y)$ .

If  $\sigma, \beta \leq 1$ , we can choose  $\phi'(x) = x^\sigma$  on  $[0, 1]$  and  $g'(x) = x^\beta$ . If  $\beta \geq \sigma \vee 1$ , then choose  $\phi' \in \mathcal{H}^\sigma(\mathbb{R})$  such that  $\phi'|_{[0,1]} \notin \cup_{b>\sigma} \mathcal{H}^b([0,1])$  and  $g'(x) = x$ . If now,  $\sigma \geq \beta \vee 1$ , we choose  $\phi' \in$

$\mathcal{H}^\sigma(\mathbb{R})$  such that  $\phi'|_{[0,1]} \notin \cup_{b>\sigma} \mathcal{H}^b([0,1])$  and such that  $\phi'(x) = x$  for all  $x \in [0, 1/2]$ . We then consider  $\zeta \in \mathcal{H}^\beta([0, 1]) \setminus \cup_{b>\beta} \mathcal{H}^b([0, 1])$  and  $g'(x) = (\zeta(x) - \zeta(0))/(2 \sup_{y \in [0,1]} |\zeta(y) - \zeta(0)|)$ .  $\square$

**6.11. Proof of Corollary 5.** Throughout this proof,  $d_\infty$  stands for the distance associated to the supremum norm  $\|\cdot\|_\infty$ . Let us define

$$\begin{aligned} \forall x, y, z \in [0, 1], \quad u(x, y) &= (u_1(x, y), u_2(x, y), u_3(x, y)) = \left( \frac{y - v_1(x)}{1 + \|v_1\|_\infty}, \frac{v_2(x)}{\|v_2\|_\infty}, \frac{v_3(x)}{\|v_3\|_\infty} \right) \\ \Phi(x, y, z) &= \|v_3\|_\infty z \varphi((1 + \|v_1\|_\infty)\|v_2\|_\infty xy). \end{aligned}$$

Let  $\mathbb{W}$  be the family of linear spaces given by Lemma 8. Define, for all  $W \in \mathbb{W}$  the linear spaces

$$T_W = \{(x, y) \mapsto a(y - f(x)), a \in \mathbb{R}, f \in W\} \quad \text{and} \quad F_W = \{(x, y, z) \mapsto z f(xy), f \in W\}$$

and set  $\mathbb{T}_1 = \{T_W, W \in \mathbb{W}\}$ ,  $\mathbb{T}_2 = \mathbb{W}$ ,  $\mathbb{T}_3 = \mathbb{W}$ ,  $\mathbb{F} = \{F_W, W \in \mathbb{W}\}$ .

It ensues from Corollary 1 of Baraud and Birgé (2011) and our Theorem 5 that there exists an estimator  $\hat{s}$  such that

$$\begin{aligned} C\mathbb{E}[H^2(s, \hat{s})] &\leq \inf_{T \in \mathbb{T}_1} \left\{ \|v_3\|_\infty^2 (1 + \|v_1\|_\infty)^{2(\sigma \wedge 1)} \|v_2\|_\infty^{2(\sigma \wedge 1)} |\varphi|_{\infty, \sigma}^2 d_2^{2(\sigma \wedge 1)}(u_1, T) + (\dim T) \tau_n^{(1)} \right\} \\ &\quad + \inf_{T \in \mathbb{T}_2} \left\{ \|v_3\|_\infty^2 (1 + \|v_1\|_\infty)^{2(\sigma \wedge 1)} \|v_2\|_\infty^{2(\sigma \wedge 1)} |\varphi|_{\infty, \sigma}^2 d_2^{2(1 \wedge \sigma)}(u_2, T) + (\dim T) \tau_n^{(2)} \right\} \\ &\quad + \inf_{T \in \mathbb{T}_3} \left\{ \|v_3\|_\infty^2 \|\varphi\|_\infty^2 d_2^2(u_3, T) + (\dim T) \tau_n^{(3)} \right\} \\ &\quad + \inf_{F \in \mathbb{F}} \left\{ d_\infty^2(\Phi, F) + (\dim F) \frac{\log n}{n} \right\} \end{aligned}$$

where

$$\begin{aligned} \tau_n^{(1)} &= \left( \log n \vee \log \left( \|v_3\|_\infty^2 (1 + \|v_1\|_\infty)^{2(\sigma \wedge 1)} |\varphi|_{\infty, \sigma}^2 \|v_2\|_\infty^{2(\sigma \wedge 1)} \right) \right) \frac{\log n}{n} \\ \tau_n^{(2)} &= \left( \log n \vee \log \left( \|v_3\|_\infty^2 (1 + \|v_1\|_\infty)^{2(\sigma \wedge 1)} |\varphi|_{\infty, \sigma}^2 \|v_2\|_\infty^{2(\sigma \wedge 1)} \right) \right) \frac{\log n}{n} \\ \tau_n^{(3)} &= \left( \log n \vee \log (\|v_3\|_\infty^2 \|\varphi\|_\infty^2) \right) \frac{\log n}{n}. \end{aligned}$$

Hence,

$$\begin{aligned} C'\mathbb{E}[H^2(s, \hat{s})] &\leq \inf_{W \in \mathbb{W}} \left\{ \|v_3\|_\infty^2 |\varphi|_{\infty, \sigma}^2 \|v_2\|_\infty^{2(\sigma \wedge 1)} d_2^{2(\sigma \wedge 1)}(v_1, W) + (\dim W) \tau_n^{(1)} \right\} \\ &\quad + \inf_{W \in \mathbb{W}} \left\{ \|v_3\|_\infty^2 (1 + \|v_1\|_\infty)^{2(\sigma \wedge 1)} |\varphi|_{\infty, \sigma}^2 d_2^{2(1 \wedge \sigma)}(v_2, W) + (\dim W) \tau_n^{(2)} \right\} \\ &\quad + \inf_{W \in \mathbb{W}} \left\{ \|\varphi\|_\infty^2 d_2^{2(1 \wedge \sigma)}(v_3, W) + (\dim W) \tau_n^{(3)} \right\} \\ &\quad + \inf_{W \in \mathbb{W}} \left\{ \|v_3\|_\infty^2 d_\infty^2(\varphi, W) + (\dim W) \frac{\log n}{n} \right\}. \end{aligned}$$

Calculating these minimums via Lemma 8 leads to the result.  $\square$

**6.12. Proof of Lemma 2.** The first part of the lemma can be deduced from Proposition 4 of Baraud and Birgé (2011). For the second part, remark that, as in the proof of Lemma 1 the problem amounts to finding  $\phi' \in \mathcal{H}^\sigma(\mathbb{R})$  with  $\phi'|_{[0,1]} \notin \cup_{a>\sigma} \mathcal{H}^a(\mathbb{R})$ ,  $v'_i \in \mathcal{H}^{\beta_i}([0,1])$  for  $i \in \{1, 2\}$ ,  $v'_1(0) = 0$ ,  $v'_2(0) = 1$  such that

$$\sqrt{v'_2} \phi'(v'_1 v'_2) \in \mathcal{H}^{\theta(\beta_1, \beta_2, \sigma)}([0,1]) \setminus \bigcup_{b>\theta(\beta_1, \beta_2, \sigma)} \mathcal{H}^b([0,1]).$$

If  $\theta(\beta_1, \beta_2, \sigma) = 2^{-1}(\beta_2 \wedge 1)$ , choose  $v'_2(x) = (1-x)^{1 \wedge \beta_2}$  and take  $\phi'$  as being any function of  $\mathcal{H}^\sigma(\mathbb{R})$  such that  $\phi'|_{[0,1]} \notin \cup_{a>\sigma} \mathcal{H}^a(\mathbb{R})$  and such that  $\phi'(0) = 1$ . If  $\theta(\beta_1, \beta_2, \sigma) = \sigma$ , choose  $v'_1(x) = 2(\sqrt{1+x}-1)$ ,  $v'_2(x) = 1/2(\sqrt{1+x}+1)$  and take  $\phi'$  as being any function of  $\mathcal{H}^\sigma(\mathbb{R})$  such that  $\phi'|_{[0,1]} \notin \cup_{a>\sigma} \mathcal{H}^a(\mathbb{R})$ . If  $\theta(\beta_1, \beta_2, \sigma) = \sigma\beta_1$ , we may assume that  $\sigma \leq 1$  and  $\beta_1 \leq 1$ . We can then choose  $v'_1(x) = x^{\beta_1}$ ,  $v'_2(x) = 1$  and  $\phi'(x) = x^\sigma$  for  $x \in [0,1]$ . If  $\theta(\beta_1, \beta_2, \sigma) = \sigma\beta_2$ , we may assume that  $\sigma \leq 1$  and  $\beta_2 \leq 1$  and choose  $v'_1(x) = 1$  for  $x \in [1/2, 1]$ ,  $v'_2(x) = 1 - (1-x)^{\beta_2}$  for  $x \in [1/2, 1]$  and  $\phi'(x) = (1-x)^\sigma$  for  $x \in [0,1]$ . Finally, if  $\theta(\beta_1, \beta_2, \sigma) = \beta_1$ , we may assume that  $\beta_1 \leq 1$ . We can then choose  $v'_1(x) = x^{\beta_1}$ ,  $v'_2(x) = (1-x)^{1 \wedge \beta_2}$  and  $\phi'$  such that  $\phi'(x) = x$  for  $x \in [0, 1/2]$ .  $\square$

**6.13. Proof of Proposition 6.** We proceed in 3 steps.

Step 1. We associate to each cube  $K \in \cup_{m \in \mathcal{M}_\ell} m$ , a place in the computer's memory. Then, for each  $i \in \{1, \dots, n\}$  we determine the sets  $K \in \cup_{m \in \mathcal{M}_\ell} m$  such that  $\mathbb{1}_K(X_i, X_{i+1}) > 0$ . There are at most  $\ell$  such sets. This permits to store all the  $\sum_{i=0}^{n-1} \mathbb{1}_K(X_i, X_{i+1})$  in around  $\mathcal{O}(n\ell d)$  operations. Let for all  $K \in \cup_{m \in \mathcal{M}_\ell} m$ ,  $I_K$  and  $J_K$  be the subsets of  $[0,1]^d$  such that  $K = I_K \times J_K$ . We can store all the  $\mu(J_K)$  in  $\mathcal{O}(4^{\ell d})$  operations and all the  $\sum_{i=0}^{n-1} \mathbb{1}_{I_K}(X_i)$  in  $\mathcal{O}(n\ell d)$  operations. This permits us to store quickly

$$\sum_{i=0}^{n-1} \mathbb{1}_K(X_i, X_{i+1}) \quad \text{and} \quad \sum_{i=0}^{n-1} \int_{[0,1]^d} \mathbb{1}_K(X_i, x) d\mu(x)$$

for all  $K \in \cup_{m \in \mathcal{M}_\ell} m$ . These values have to be calculated to know the  $F_K(K')$  and thus to use the algorithm presented in Section 5.

Step 2. For each  $K \in \cup_{m \in \mathcal{M}_\ell} m$ , we use the algorithm of Section 5 to design  $m'_K$ . Let us denote by  $j \in \{0, \dots, \ell\}$  the smallest integer such that  $K \in \mathcal{K}_j$  where  $\mathcal{K}_j$  is defined in Section 2.2.

- To find  $m'_K$ , we begin to compute  $\mathcal{E}(T^*(K''))$  for all  $K'' \in \cup_{m \in \mathcal{M}_\ell \setminus \mathcal{M}_{\ell-1}} m$  such that  $K'' \cap K \neq \emptyset$ . The complexity of this is around the number of such sets, i.e.,  $4^{(\ell-j)d}$ .
- Next, thanks to relation (12) we compute  $\mathcal{E}(T^*(K''))$  for all  $K'' \in \cup_{m \in \mathcal{M}_{\ell-1} \setminus \mathcal{M}_{\ell-2}} m$  such that  $K'' \cap K \neq \emptyset$ . There are  $4^{(\ell-j-1)d}$  such sets. The complexity of this operation is thus  $4^d \times 4^{(\ell-j-1)d}$ .
- By recurrence, we compute  $\mathcal{E}(T^*(K''))$  for all  $K'' \in \cup_{m \in \mathcal{M}_\ell \setminus \mathcal{M}_j} m$  such that  $K'' \cap K \neq \emptyset$  in at most

$$4^{(\ell-j)d} + 4^d \times \sum_{k=1}^{\ell-j-1} 4^{kd} \leq 3 \times 4^{(\ell-j)d}$$

operations.

- We get then  $\mathcal{E}(T^*([0, 1]^d))$  in  $4^d j$  additional operations.

We apply this algorithm for all  $K \in \cup_{m \in \mathcal{M}_\ell} m$ . When  $K \in \mathcal{K}_j$ , computing  $m'_K$  requires thus  $\mathcal{O}(4^{(\ell-j)d} + 4^d j)$  operations. Since  $|\mathcal{K}_j| = 4^{jd}$ , computing all the  $m'_K$  requires finally

$$\sum_{j=0}^{\ell} 4^{jd} (4^{(\ell-j)d} + 4^d j) = \mathcal{O}(\ell 4^{(\ell+1)d})$$

operations.

Step 3. Now, by slightly modifying the algorithm, we can compute (10) in  $\mathcal{O}(4^{(\ell+1)d})$  operations.

□

Acknowledgements: many thanks to Yannick Baraud for his suggestions, comments, careful reading of the paper. We are thankful to Claire Lacour for sending us the source code of the procedure of Akakpo and Lacour (2011).

## REFERENCES

- Akakpo, N. (2009). *Estimation adaptative par sélection de partitions en rectangles dyadiques*. PhD thesis, Université Paris Sud.
- Akakpo, N. (2012). Adaptation to anisotropy and inhomogeneity via dyadic piecewise polynomial selection. *Mathematical Methods of Statistics*, 21:1–28.
- Akakpo, N. and Lacour, C. (2011). Inhomogeneous and anisotropic conditional density estimation from dependent data. *Electronic Journal of Statistics*, 5:1618–1653.
- Athreya, K. B. and Atuncar, G. S. (1998). Kernel estimation for real-valued Markov chains. *Sankhyā. The Indian Journal of Statistics. Series A*, 60:1–17.
- Baraud, Y. (2010). Estimator selection with respect to hellinger-type risks. *Probability Theory and Related Fields*, pages 1–49.
- Baraud, Y. and Birgé, L. (2009). Estimating the intensity of a random measure by histogram type estimators. *Probability Theory and Related Fields*, 143:239–284.
- Baraud, Y. and Birgé, L. (2011). Estimating composite functions by model selection. *Annales de l'Institut Henri Poincaré (B) Probability and Statistics*. To appear.
- Basu, A. K. and Sahoo, D. K. (1998). On Berry-Esseen theorem for nonparametric density estimation in Markov sequences. *Bull. Inform. Cybernet.*, 30(1):25–39.
- Birgé, L. (1983). Approximation dans les espaces métriques et théorie de l'estimation. *Probability Theory and Related Fields*, 65:181–237.
- Birgé, L. (1984a). Stabilité et instabilité du risque minimax pour des variables indépendantes équidistribuées. *Annales de l'Institut Henri Poincaré. Probabilités et Statistique*, 20:201–223.
- Birgé, L. (1984b). Sur un théorème de minimax et son application aux tests. *Probability and Mathematical Statistics*, 2:259–282.
- Birgé, L. (2006). Model selection via testing: an alternative to (penalized) maximum likelihood estimators. *Annales de l'Institut Henri Poincaré (B) Probability and Statistics*, 42(3):273 – 325.
- Birgé, L. (2007). Model selection for poisson processes. *Lecture Notes-Monograph Series*, 55:pp. 32–64.
- Birgé, L. (2008). Model selection for density estimation with  $\mathbb{L}_2$ -loss. *ArXiv e-prints*.

- Birgé, L. (2012). Robust tests for model selection. *Festschrift for Jon Wellner, IMS Collections*. To appear.
- Blanchard, G., Schäfer, C., and Rozenholc, Y. (2004). Oracle bounds and exact algorithm for dyadic classification trees. 3120:378–392.
- Bradley, R. C. (2005). Basic properties of strong mixing conditions. A survey and some open questions. *Probability Surveys*, 2:107–144.
- Cléménçon, S. (2000). Adaptive estimation of the transition density of a regular Markov chain. *Mathematical Methods of Statistics*, 9:323–357.
- Comte, F. and Rozenholc, Y. (2002). Adaptive estimation of mean and volatility functions in (auto-)regressive models. *Stochastic Process. Appl.*, 97(1):111–145.
- Dahmen, W., DeVore, R., and Scherer, K. (1980). Multi-dimensional spline approximation. *SIAM Journal on Numerical Analysis*, 17(3):pp. 380–402.
- DeVore, R. and Yu, X. (1990). Degree of adaptive approximation. *Mathematics of Computation*, 55:625–635.
- Dorea, C. C. Y. (2002). Strong consistency of kernel estimators for Markov transition densities. *Bull. Braz. Math. Soc. (N.S.)*, 33(3):409–418. Fifth Brazilian School in Probability (Ubatuba, 2001).
- Doukhan, P. (1994). *Mixing*, volume 85 of *Lecture Notes in Statistics*. Springer-Verlag. Properties and examples.
- Doukhan, P. and Ghindès, M. (1983). Estimation de la transition de probabilité d’une chaîne de Markov Doëblin-récurrente. Étude du cas du processus autorégressif général d’ordre 1. *Stochastic Process. Appl.*, 15(3):271–293.
- Hochmuth, R. (2002). Wavelet characterizations for anisotropic besov spaces. *Applied and Computational Harmonic Analysis*, 12:179 – 208.
- Juditsky, A., Lepski, O., and Tsybakov, A. (2009). Nonparametric estimation of composite functions. *Ann. Stat.*, 37(3):1360–1404.
- Lacour, C. (2007). Adaptive estimation of the transition density of a Markov chain. *Annales de l’Institut Henri Poincaré. Probabilités et Statistiques*, 43:571–597.
- Lacour, C. (2008). Nonparametric estimation of the stationary density and the transition density of a Markov chain. *Stochastic Processes and their Applications*, 118:232–260.
- LeCam, L. (1973). Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, 1:38–53.
- LeCam, L. (1975). *On local and global properties in the theory of asymptotic normality of experiments*. Academic Press.
- Massart, P. (2003). *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer Berlin/Heidelberg. École d’été de Probabilités de Saint-Flour.
- Roussas, G. (1969). Nonparametric estimation in Markov processes. *Annals of the Institute of Statistical Mathematics*, 21:73–87.
- Roussas, G. G. (1991). *Estimation of transition distribution function and its quantiles in Markov processes: strong consistency and asymptotic normality*, volume 335 of *NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci.* Kluwer Acad. Publ., Dordrecht.
- Sart, M. (2012). Model selection for poisson processes with covariates. *ArXiv e-prints*.
- Viennet, G. (1997). Inequalities for absolutely regular sequences: application to density estimation. *Probability Theory and Related Fields*, 107:467–492.



UNIVERSITÉ DE NICE SOPHIA-ANTIPOLIS, LABORATOIRE J-A DIEUDONNÉ, PARC VALROSE, 06108 NICE CEDEX 02, FRANCE.

*E-mail address:* `msart@unice.fr`